Teaching Literacy – A Holistic Reframing of Teacher Assessment Literacy

Charles A. DePascale

Psychometric Confections, LLC

April 2021

Prepared for the 2021 annual meeting of the New England Educational Research Organization, April 28-30, 2021

Teaching Literacy – A Holistic Reframing of Teacher Assessment Literacy

There has been longstanding and widespread agreement that some degree of assessment literacy is an essential component of effective instruction. It is also universally acknowledged that efforts to enhance teachers' assessment literacy historically have been inadequate and largely unsuccessful. As the concept of assessment literacy has evolved, recent efforts are much more focused on the use of assessment by teachers within the context of instruction. To a certain extent, however, there is a lingering perspective that assessment is important to support instruction, but is different or separate from instruction. At the same time, the terms instruction and teaching have become synonymous. In this paper, I propose a more holistic perspective of assessment as an inseparable component of teaching, that is, there can be no teaching without assessment. From that perspective, teacher assessment literacy can best be viewed as teaching literacy. In the final sections of the paper, implications for supporting teachers' interpretation and use of external, large-scale test results are discussed along with recommendations for reporting results from large-scale tests in a way that supports teaching literacy.

Assessment literacy is critical for effective teaching and K-12 teachers do not have enough of it. On those two points there appears to be longstanding and widespread agreement. There also appears to be growing consensus on the definition of assessment literacy and a coalescing of opinions regarding the knowledge, skills, and abilities (or competencies, if you prefer) that it comprises. There remains much less agreement, however, on how best to improve teachers' assessment literacy; and scalable approaches or specific programs that have been effective in increasing teachers' assessment literacy continue to elude us.

As an educational assessment and measurement community, we have attacked the problem of teachers' assessment literacy from our perspective. We have identified key assessment and measurement concepts that teachers *need to know* and have generated countless books, articles, instructional modules, and other media in an attempt to convey those key concepts to teachers. To date, that body of information has had limited impact on enhancing teachers' assessment literacy or on increasing the productive use of assessment to inform teaching and learning. If the goal is to make an impact on teachers' assessment literacy, it seems that a new approach is necessary, an approach that begins with empathy and understanding on our part of teachers' perspectives on assessment and also includes a more holistic view of the role that assessment plays in teaching.

Barriers to Enhancing Teachers' Assessment Literacy

Research has documented several factors that are likely contributors to the ongoing difficulty in increasing teachers' assessment literacy. Historically, chief among these is the dosage of instruction teachers receive in assessment literacy; that is, the amount of training teachers receive in assessment, when the training occurs, and the frequency with which it occurs is insufficient to produce and sustain necessary levels of assessment literacy (Gullickson and Ellwein, 1985; DePascale, 1989; Popham, 2011; DeLuca and Bellara, 2013; Greenberg, McKee, and Walsh, 2013). DeLuca (2012) also identified the treatment of assessment literacy as a set of assessment-related knowledge and skills isolated from the context of teaching as another barrier. Xu and Brown (2016) expanded thinking on the barriers to increasing teachers' assessment literacy by reconceptualizing assessment literacy as teacher assessment literacy in practice (TALiP). Their TALiP framework describes three levels of mastery necessary for teacher assessment literacy:

- The first level is the traditional concept of mastery of education assessment knowledge.
- Second is an internalized set of understanding and skills of the interconnectedness of assessment, teaching, and learning.
- Third is self-directed awareness of assessment processes and one's own identify as an assessor. (p. 159)

Central to the TALiP framework is teachers' ability to acknowledge and accept their identity as assessor along with their identify as teacher. As Xu and Brown state, "according to the proposed framework, assessment literate teachers are those who constantly reflect on their assessment practice, participate in professional conversations about assessment in communities, engage in professional conversations about assessment, self-interrogate their conceptions of assessment, and seek resources for to gain a renewed understanding of assessment and their own roles as assessors" (p. 159).

In this paper, I attempt to advance thinking on teachers' assessment literacy by taking the TALiP

framework's conception of teacher as assessor one step further by fully encompassing assessment within a teacher's identity as a teacher. That is, I posit that teacher's assessment literacy can best be enhanced by moving beyond the current practice of thinking of assessor as distinct from teacher or in thinking of assessment as a separable component of teaching and to regard it as a sine qua non of teaching; that is, there is no teaching without assessment and therefore, assessment literacy is best viewed as simply a part of *teaching literacy*.

From the perspective of the teacher, this framing of assessment literacy a) moves the conversation beyond the long-ago dispelled belief that teaching can occur without assessment, and b) reduces problems associated with asking teachers to embrace multiple identities (e.g., instructor, assessor, psychologist, entertainer, comforter, arbitrator, manager, disciplinarian). From our perspective as assessment specialists hoping to support teachers' assessment literacy, this framing of assessment literacy as teaching literacy compels us to view assessment as teachers use and understand assessment rather than simply sitting on the outside looking in and asking teachers to better understand how to take what we know about assessment and apply it to their work.

New Premise – teacher as teacher

In *Building A Conceptual Framework for Assessment Literacy* (DePascale, Sharp, Ryan, and Betebenner, 2018), my colleagues and I set out to develop a deeper understanding of assessment literacy with the ultimate goal of improving the design of instructional and professional development modules intended to improve the assessment literacy of key stakeholders in K-12 education, teachers, school administrators, and policymakers. In the course of reviewing and synthesizing the research on assessment literacy for that paper, we reached three conclusions:

- Assessment literacy is multidimensional requiring a degree of understanding of principles relevant to data literacy, measurement, and test development.
- The amount and blend of those data, measurement, and test development skills needed to be considered assessment literate is context dependent and varies significantly for teachers, school administrators, and district or state policymakers.
- Assessment literacy is an applied concept completely nested within the context in which it is being considered.

It is that third conclusion that leads now to the more holistic view of teacher's assessment literacy as *teaching literacy*. In this context, I am using the definition of holistic as the "comprehension of the parts of something as intimately interconnected and explicable only by reference to the whole." In considering the research on teachers' assessment literacy, we concluded that it is impossible to consider teachers' assessment literacy outside of the context in which teachers use assessment as part of the teaching process. (Brookart, 2011; DeLuca, 2012; Xu and Brown, 2016).

Teachers use of assessment information

Ideally, teachers use assessment information to inform decision making at various stages of the teaching process. The manner in which they use assessment and the effectiveness with which they use it, however, are dependent upon their ability to know what to ask, whom to ask, when to ask, and what to do with the information once they once they have assessed.

This last point has been a major sticking point in the design of instructional support and assessment systems for decades. There is a desire, or even an expectation, that the assessment instrument provides information to guide what the teacher should do next. The usefulness and need for that type of information is real and obvious. What has been less obvious, particularly to assessment and measurement specialists, however, is that an assessment instrument alone cannot be the source of information on what a teacher should do next. Perhaps more importantly, assessment specialists and psychometricians lack the knowledge base needed to provide that type of information on the basis of a test score or student performance on an assessment task. Information about *what do next* is deeply embedded within the instructional context and is dependent on a teacher's ability to process the data or information collected via assessment along with other relevant information that defines the specific context.

A teacher's use of student responses during an assessment activity to inform instructional decisions lies at the higher end of the "data, information, knowledge, wisdom" pyramid (DIKW). Whether you ascribe to the DIKW hierarchy or alternate conceptions of knowledge and knowledge management, there is general agreement that those with the highest-level skills not only possess information or knowledge, but are able to "use that practical know how to achieve appropriate ends." (Frické, 2007).

It is impossible, therefore, to separate the concept of assessment literacy as the "small number of terribly significant understandings… that a person has about the fundamental procedures and concepts in educational measurement that are apt to influence educational decisions" (Popham, 2018) from the knowledge or wisdom necessary for a teacher to apply the information gained from assessment in an instructional setting. That knowledge or wisdom lies completely outside of the realm of assessment or measurement theory. Attempts, therefore, to frame assessment literacy from an assessment/measurement perspective rather than a teaching perspective are destined to fail, or at best, have limited success.

A Teacher's Way of Knowing

The concept that teachers' knowledge of a particular subject area or domain is unique or peculiar to teachers and teaching is at least as old as public education in the United States. Mann (1840) in describing the aptness to teach noted "the ability to acquire and the ability to impart are wholly different talents." He further described the aptness to teach in terms of the following skills:

- The power of discovering and solving at the time the exact difficulty by which the learner is embarrassed.
- The removal of a slight impediment, the drawing aside of the thinnest veil which happens to divert his steps or obscure his vision.
- How much does the pupil comprehend of the subject? What should his next step be? Is his mind looking toward a truth or an error? The answers to these questions must be intuitive in the person who is apt to teach.

A century later, the work of Shulman and his colleagues captured the essence of Mann's description a teacher's knowledge and laid the foundation for a new body of educational research on the concept of *pedagogical content knowledge*. As described by Shulman (1986), *pedagogical content knowledge* is "a type of knowledge that is unique to teachers, and is based on the manner in which teachers related their pedagogical knowledge (what they know about teaching) to their subject matter knowledge (what they know about what they teach). Like Mann's aptness to teach, teachers with pedagogical content knowledge possess

- ways of representing and formulating the subject that makes it comprehensible to others,
- an understanding of what makes the learning of specific concepts easy or difficult: the conceptions and preconceptions that students of different ages and backgrounds bring with them to the learning, and
- must have at hand a veritable armamentarium of alternative forms of representation.

Shulman's original concept of pedagogical content knowledge as a synthesis of pedagogical knowledge and content knowledge (i.e., knowledge of subject matter, discipline) was expanded by Cochran, DeRuiter, and King (1993) who extended the concept to include two additional components:

- teachers' knowledge of students' abilities and learning strategies, ages and developmental levels, attitudes, motivations, and prior knowledge of the concepts to be taught, and
- teachers' understanding of the social, political, cultural and physical environments in which students are asked to learn.

The addition of components related to students and the context in which instruction occurs is consistent with the evolution of the concept of teacher and teaching since the last quarter of the twentieth century. The ability of effective teachers to blend and apply knowledge in these four areas: pedagogy, subject matter, student, and social context, supports the aphorism that Shulman repurposed and posited in the title of his 1986 article, those who understand, teach.

Research over the decades that followed further refined the meaning of pedagogical content knowledge within specific subject areas, age ranges or levels of education, and documented the development of pedagogical content knowledge within teachers (Gudmundsdottir and Shulman, 1987; Magnusson, Krajcik, and Borko, 1999; Flood, Lapp, Squire, and Jensen, 2003; Hill, Schilling, and Ball, 2004; Heritage and Vendlinski, 2006; Hill, Ball, and Schilling, 2008; Shulman and Shulman, 2008,2009). All of this research helps to firmly establish that decision making is at the core of teaching and that to a large extent those decisions are context dependent.

Teaching, Decision Making, Assessment

Educational assessment at its core is a process for gathering evidence to inform decision making that ideally will lead to an improvement in student learning. That evidence might be related to decisions made by a teacher and a third-grade student

regarding the performance of that individual student, by a teacher regarding the performance of a third-grade class, by a school principal regarding the performance of third grade students and teachers within their school or a particular intervention's effectiveness for those third grade students and teachers, up to decisions made by a commissioner of education and department of education related to the performance of third grade students in their state. The evidence collected and the instruments used to collect it will vary appropriately for each of those decision-making contexts. In this paper, I am concerned specifically with decisions made by teachers and the ways in which they use assessment (i.e., the process of assessment) to gather evidence to inform those decisions.

When teaching is viewed in terms of decision making to inform and improve interactions with students, the role of assessment as an integral, integrated component within the teaching process becomes clearer. Decision making should be supported by evidence. Evidence, both informally and formally collected, comes primarily from the assessment process. In the context of the instructional process, most evidence to inform instructional decision making will be derived from formative assessment processes.

Describing the knowledge that teachers need to successfully employ formative assessment processes during instruction, Heritage (2007) includes knowledge "about the range of formative assessment strategies so that they can maximize the opportunities for gathering evidence" (p. 143). Among the skills needed by teachers to effectively use formative assessment Heritage includes: "drawing inferences from students' responses…no matter what the assessment strategy – observation, dialogue, asking for a demonstration or a written response – teachers must examine students' responses from the perspective of what they show about their conceptions, misconceptions, skills, and knowledge" (p. 144). The ultimate goal of this process being "to infer what the 'just right gap' is between current learning and desired goals, identifying students emerging understanding or skills so that they can build on these by modifying instruction to facilitate growth" p. (144). Implicit, but not mentioned in this paragraph, are the other knowledge and skills (i.e., pedagogical content knowledge) that a teacher must possess in order to make the correct inferences, identify the gap, and modify instruction, if necessary.

The interconnectedness of instruction and assessment that is self-evident in the formative assessment process is not a new concept. Every student of education has encountered some version of the ubiquitous curriculum-instruction-assessment triangle early in their coursework. With curriculum, instruction, and assessment as its vertices, the interior of the triangle is often depicted with a statement related to students: how they learn (e.g., the learning process, learning theory, teaching) or what they learn (e.g., state content standards) to further reflect the interdependency of the three components. Recent work by Shepard and colleagues emphasizes the importance of learning theory in the process and explicitly connects assessment and instruction, describing formative assessment as practices that intersect with those intended to support deep learning and participation in disciplinary discourse practices. (Penuel and Shepard, 2016; Shepard, Penuel, Davidson, 2017; Shepard, 2019). Starting from this perspective, one might expect that the preparation of teachers, curriculum specialists, assessment specialists, etc. would recognize and respect the interdependence. That, however, has not been the case, at least in the United States.

The Isolation of Curriculum, Instruction, and Assessment

Many factors may have contributed to the separation and isolation of curriculum, instruction, and assessment. The valuing of specialization over generalization in higher education certainly plays some role in the ongoing development of curriculum, instruction, and assessment as three distinct academic disciplines. The emergence of curriculum, instruction, and assessment as distinct disciplines alone, however, was not sufficient to cause the separation between teaching and assessment that we are now trying to address. The introduction of the scientific method into education at the beginning of the twentieth century and along with it the *cult of efficiency* that helped shape public education in the United States in response to the Industrial Revolution and waves of immigration likely also played a role, influencing views on curriculum, teaching and assessment for much of the twentieth century (Bobbit, 1918; Callahan, 1964). As described by Kliebard (2004), "it was social efficiency that, for most people, held out the promise of social stability in the face of cries for massive social change." (p. 76). With regard to teaching, the emphasis on efficiency resulted in a focus on teacher and student behavior management metrics as indicators of effective teaching. That teacher-centered, management-focused view defined and dominated research on teaching until the last quarter of the century. With regard to assessment, the focus shifted from the

classroom to external, standardized tests used to inform decisions outside of the teaching process.

It may be that defining teaching in terms of efficiency and teacher-centered behaviors rather than in terms of learning-centered interactions between teachers and students effectively cut teaching from the whole curriculum-instruction-assessment triangle as much as it separated the three components from each other or in some way isolated instruction from assessment. Having separated teaching from curriculum-instruction-assessment, when interest in teaching expanded from efficiency to include teacher effectiveness, it became much easier to focus on things that the teacher does (processes and procedures) rather than on the interactions and decision-making events involving students.

From there, it is an easy logical step to arrive at definitions and descriptions of curriculum, instruction, and assessment as three distinct components with instruction alone being synonymous with teaching.

- Curriculum described as providing the details on what will or should be taught.
- Instruction described as the methods and actions used to implement the curriculum with students to improve student learning.
- Assessment described as the process of, or more often the instruments used to, measure whether students have learned what has been taught and/or what was supposed to be taught.

There is a clear temporal sequence to the three components when curriculum, instruction, and assessment are defined in this manner. Curriculum precedes instruction. Assessment follows instruction. Even if we allow for a feedback loop (i.e., assessment can inform instruction) the mindset is established that assessment occurs after instruction – no matter how close the proximity between instruction and assessment.

Reframing Assessment Literacy from a Teacher's Perspective

If the goal is to reframe assessment literacy from a teacher's perspective based on a holistic view of curriculum, instruction, and assessment as inseparable elements of teaching, one of the first questions to consider is who should be involved in the process and who should not be involved. The answer to who should not be involved, or should play a much more limited role than they currently do, seems most obvious. Large-scale assessment specialists have attempted to play a leading role in developing assessment literacy protocols and programs, but their efforts have seen limited success. Under this reframed paradigm, it seems logical that their role should be limited to areas where they are best suited to provide support; that is, focusing on issues related to the use of large-scale tests to support education policy rather than teachers' decision-making – acknowledging the limited role that large-scale tests can and should play in teachers' decision-making. Similarly, the National Council on Measurement in Education (NCME), which has shown a recent interest in classroom assessment and enhancing teachers' assessment literacy, can only be expected to play a limited role in promoting teachers' assessment literacy given that the primary focus of the organization has not been on teachers and teaching. Those within the organization who have focused on teaching and assessment can best support the organization and teachers' assessment literacy by serving as a filtered conduit between measurement professionals and the classroom – ensuring that necessary and appropriate information flows in both directions.

Those who should be involved are the people who have been actively involved in promoting the use of assessment for teaching and learning the past several decades, that is, those who have played a key role in promote formative assessment as a critical part of teaching.

Applying Educational Measurement Principles to Teaching

Much of the work in assessment literacy has been devoted to increasing teachers' understanding of the fundamental principles of educational measurement and assessment. The starting points, of course, for all discussions of educational testing are validity, reliability, and fairness. What do those principles look like when applied to teaching and the use of assessment by teachers to inform their decision making?

Historically, those principles were defined almost exclusively in terms of large-scale testing. More recently, much more attention has been focused on the application of those principles to formative assessment and other types of classroom-based assessment. Some have suggested that the metrics remain unchanged when translating these principles from large-scale testing to the classroom, but the criteria or benchmarks for what is acceptable vary based on differences in the consequences associated with any single assessment or decision. Others have suggested that the field needs to rethink and reframe

the principles to make them more appropriate for the classroom setting (Brookhart, 2003; Moss, 2003; Kane and Wools, 2019; Shepard, 2019).

There is merit to both viewpoints when educational measurement and assessment is viewed from the perspective of large-scale testing. The applications of validity, reliability, and fairness that have evolved for large-scale testing would have to be adapted for classroom testing and would certainly have to be adapted to be applied to teaching. The problem with this perspective, however, is that it is focused on the application of validity, reliability, and fairness to a particular context and often to a particular assessment instrument (i.e., a large-scale test, a classroom test) rather than focusing on the essence of those principles as they relate to the process of assessment and measurement, of which tests are simply a tool. When one considers the basic definitions of validity, reliability, and fairness and their purpose in the educational assessment context, the concepts apply equally well and are equally relevant to the teaching, large-scale assessment, and classroom assessment contexts. The specifics of how they are applied, of course, should vary, as appropriate, for the particular context.

Validity

It is firmly established within the educational measurement community that validity is concerned with the meaning of test scores and the inferences drawn from test results and is not an attribute of the tests or the test scores themselves. The definition of validity in terms of whether an inference, or claim, is supported for a particular purpose can easily be transferred to teaching and the use of assessment to support teachers' decision making. The definition of validity itself is test- or instrument-free. The quality and characteristics of the test instruments come into play only in the context of whether a specific instrument provides suitable evidence to support the inferences that must be made.

The specifics of how the principles of validity are applied and evaluated will vary based on the questions being addressed and the design of the instrument(s) used to collect evidence, but the general elements of validity contained in Messick's Facets of Validity as Progressive Matrix apply to the teaching context as well as they do to a large-scale testing context: construct validity, construct validity + value implications, construct validity + relevance/utility, and construct validity + relevance/utility + value implications + social consequences. As described by Messick (1990), the four elements "correspond to four interrelated aspects of the basic validity question: "To what degree, if at all, on the basis of evidence and rationales, should the test scores be interpreted and used in the manner proposed?" It is easy to see a direct connection between evidence collected to inform decision making during teaching, particularly evidence collected via formative assessment, and the "four distinguishable but interrelated aspects" of the validity question laid out by Messick:

- what balance of evidence sustains the interpretation or meaning of the scores.
- what evidence supports not only score meaning, but also the relevance of the scores to the particular applied purpose and the utility of the scores in the applied setting.
- what makes credible the value implications of the score interpretation and any associated implications for action; and,
- what signifies the functional worth of the testing in terms of its intended and unintended consequences.

Reliability

Reliability is most often defined in terms of consistency, accompanied by a statement acknowledging the reality that no measurement is perfect. The implication is that test scores, observations, and other measurements that are more consistent are more reliable, a desirable outcome. Consistency in and of itself, however, is not the goal when considering reliability in the context of educational measurement or when attempting to develop instruments that produce more reliable scores. Rather, the goal is to achieve consistency while compiling enough of the right type of evidence to allow you to make an accurate estimate of student achievement and to make accurate inferences about what the student can and cannot do. That is, like validity, the concept of reliability is tied to the inferences that can be supported by the test scores and is not a characteristic of the test instrument itself (Traub and Rowley, 1991). Sufficiency, rather than consistency, may be the more appropriate term to describe the core characteristic of reliability.

The relevance of sufficiency to the concept of reliability is clear even within the context of large-scale tests. As large-scale testing specialists we were taught of the interconnectedness of reliability and validity. Validity is capped by reliability; that is, reliability is a necessary but insufficient condition for validity. Designing tests, however, we quickly learned of the complexity of the relationship between validity and reliability. Design decisions intended to increase

reliability (e.g., use only selected-response items, make the test content more homogeneous) could have a negative impact on validity by restricting the inferences that could be supported from the resulting test scores. The same held true when considering reliability in terms of generalizability. So-called generalizability indices can be maximized by making choices that limited the generalizability of the inferences supported by test scores. Efforts to maximize reliability cannot be made without regard their effect on optimizing validity. Similarly, efforts to maximize validity cannot be made without regard to their impact on reliability.

Moss (1994) posited that there are other approaches beyond the psychometric conception of reliability that can and should be considered with respect to evaluating the technical quality or appropriateness of less standardized forms of assessment such as performance assessments. More importantly, Moss pointed out that much of what we consider as the key characteristics of reliability in education measurement are applications of the general measurement principles that apply specifically to discrete large-scale test instruments rather than general principles in their own right that must, or even should, be applied to other forms of assessment and educational measurement (Moss, 1994, 2003). Therefore, rather than needing to reframe reliability so that it fits within the classroom setting, we need simply to step back from large-scale tests to consider the fundamental principles of reliability that led to those specific procedures and protocols.

With regard to assessment in the classroom, the concept of reliability is best considered in terms of the question, have teachers collected enough evidence and the right type of evidence to support the decision that they need to make. From this perspective, we can think of reliability in terms of whether the evidence collected is sufficient to support teachers' inferences (Brookhart, 2003; Smith, 2003).

Fairness

Fairness is arguably the least technically laden concept among the three pillars of educational measurement: validity, reliability, and fairness. In fact, the Joint Standards for Educational and Psychological Testing state, "the term *fairness* has no single technical meaning," and that "it is possible that individuals endorse fairness in testing as a desirable social goal, yet reach quite different conclusions about the fairness of a given testing program" (APA, NCME, AERA, 2014, p. 49). There is general agreement, however, that fairness, like reliability, is fundamentally a validity issue. As such, fairness is tied to the question of whether there is sufficient evidence to support the claim that the inferences that are being made to support specific decisions apply equally well to all students or groups of students.

The Joint Standards identify four broad categories of fairness in testing that apply equally well to our conception of assessment as part of the teaching process:

- fairness in treatment during the testing process
- fairness in access to the construct(s) being assessed
- fairness as validity of individual test score interpretations for the intended uses
- fairness as lack of measurement bias.

The application of the first three categories of fairness to teaching and teaching literacy should be self-evident and requires no further discussion here. The final category, fairness as lack of measurement bias, can be recast from the technical definitions and measures of bias that apply to large-scale tests to the principles that those measures are designed to enforce. The key questions being addressed with regard to measurement bias are whether a) the manner in which the student is being assessed is susceptible to factors other than the construct(s) being assessed or b) the meaning of outcomes resulting from the assessment process apply equally well to all students or groups of students. Although there are statistical tests of measurement bias applied to large-scale tests, many of the practices related to reducing or eliminating opportunities for bias on large-scale tests are based on non-technical professional judgments made during the item development and test design process. The JCSEE Classroom Assessment Standards, Standard Q3 Unbiased and Fair Assessment, provides guidelines for applying a similar review process to assessment in the classroom.

A final note about fairness, teaching, and assessment is that the increased attention to cultural awareness and cultural responsiveness in the classroom is likely to have a profound impact over the next few years on teaching, the development and use of classroom tests, and the development and use of large-scale tests. Some of the expected reforms are likely to be incremental additions or continuations of the improvements that have been ongoing over the past decade with regard to special populations, that is, students with disabilities and English learners. Other reforms may be more foundational and it will be

beneficial to consider teaching and large-scale tests jointly as changes to current practices are considered.

Building Teaching Literacy

Having reframed assessment literacy as teaching literacy, we must return to the issue raised in the opening sentence of this paper: assessment literacy is critical for effective teaching and K-12 teachers do not have enough of it. How then do we approach building teaching literacy in a way that is more successful than decades of attempts at building teachers' assessment literacy have been?

The first step, I believe, is to separate the concepts of teaching literacy, defined here in terms of issues related to teachers' decision making from large-scale assessment literacy, defined in terms of the knowledge and skills related to the design, development, interpretation, and use of large-scale tests and their results. This separation should be as clean and complete as possible so as not to conflate the broad issue of the literacy in assessment needed with respect to the design, development, administration, interpretation and use large-scale tests, in general, and the much narrower issue of the knowledge and skills that teachers might need to incorporate results from large-scale tests into their decision-making process. As discussed previously, Moss identified the problems associated with the situation where many of the educational measurement principles accepted as applying to all aspects of assessment, were in fact, applications of general principle that applied specifically to discrete, large-scale tests. The same can be said of approaches to assessment literacy that focus on statistical indicators and technical characteristics more appropriately associated with large-scale tests than those associated with teaching and assessment in the classroom. In the final section of this paper, we will address issues related to supporting teachers' use of large-scale tests.

The second step is to understand that formative assessment is inseparable from teaching (and quite separable from large-scale testing). It is important to acknowledge the work that is already being done to expand the body of research associated with formative assessment processes and to build teachers' knowledge, skills, and use of those practices to inform their decision making and students' own involvement in their learning and in their peers' learning. Pre-service and in-service efforts that focus on formative assessment as a process to support teachers' decision making are already performing a significant portion of the work needed to improve teaching literacy.

The next step in the process of building teaching literacy is to move beyond attempting to find effective ways to enhance teachers' assessment literacy and refocus energy toward ensuring that the fully developed concept of *teaching literacy* is part and parcel of all pre-service teacher preparation programs and in-service professional development opportunities for teachers. The barriers to current teachers and even prospective teachers embracing assessment as a part of their role as teacher that were identified by Xu and Brown (2016) still must be overcome, of course. Separating large-scale testing from the concept of assessment as an integral part of teachers' decision- making process is a starting point.

Acquiring and Enhancing Proficiency in Teaching Literacy

In his seminal article on pedagogical content knowledge, Shulman (1986) calls for a body of research devoted to teacher learning. Over the decades that followed, significant research has been conducted on teaching and how teachers acquire teaching skills in various content areas, with regard to areas of focus such as social justice and cultural awareness, as well as on teachers as learning individuals and as member of a community. Shulman also describes the use of case method in teacher education "as a means for developing strategic understanding, for extending capacities toward professional judgement and decision-making." Although the specifics of a applying a case method approach may have evolved since 1986, the fundamental concept of embedding training in the use of assessment within actual decision-making situations faced by teachers remains sound. Applying current learning theory (related both to teachers as teachers and teachers as students) and using a constructivist approach to designing cases, in which assessment (i.e., the gathering of evidence) to inform a teacher's decision making is the centerpiece, can be a first step in helping prospective and novice teachers acquire proficiency in teaching literacy. As teachers realize that they require more and better evidence to improve their decision making and seek ways to acquire that evidence, their teaching literacy will increase.

Teachers moving from novice to professional to expert will apply assessment practices and make decisions based on evidence gained from assessment in a manner that appears instinctive or intuitive, but that intuition will be grounded in pedagogical content knowledge and experience, and those teachers will also possess the ability to reflect on what they have done and explain why they took certain actions (Wylie and Lyon, 2016). As Shulman (1986) describes, "a

professional is capable not only of practicing and understanding his or her craft, but of communicating the reasons for professional decisions and actions to others."

Shulman (1986) also recommends a method for the development of professional examinations for teachers that is defined and controlled by members of the profession and is grounded in the content and processes needed by teaching professionals, and which "will be useful when only those who have been professionally prepared as teachers are likely to pass them because they tap the unique knowledge bases of teaching" (p. 13). Separating the assessment itself from the politics surround its current use, the edTPA is a performance assessment developed by teachers and teacher educators based on the "conviction that performance assessments in teaching are a critical strategy for developing teachers' expertise and for improving the quality of teacher preparation" (Darling-Hammond and Hyler, 2013, p. 10). The edTPA, or a similar instrument, would appear to be an attempt to fulfill the Shulman's vision of an assessment for pre-service and beginning teachers.

In summary, it appears that we have begun to assemble some of the building materials needed to support pre-service teachers' acquisition of teaching literacy. However, as Coladarci (2002) noted with respect to local assessment systems, "a collection of assessments does not entail a system any more than a pile of bricks constitutes a house" (p. 773). We need to move beyond isolated or idiosyncratic research-based interventions of formative assessment practices and the development of one highly charged assessment system for pre-service and beginning teachers to the design, development, and implementation of teacher preparation programs that foster a view of teaching that fully integrates assessment and instruction. Although some adjustment to content and philosophy will be necessary, I am optimistic that the current teacher preparation systems which have focused more heavily on the practicum experience of pre-service teachers and the induction of beginning teachers would not have to be torn down and rebuilt to produce beginning teachers who have at least a novice's level of teaching literacy.

A final piece in the puzzle is to support beginning teachers' progression from novice to professional to expert, as well as to support enhancing the teaching literacy of currently practicing teachers. Kane (1994) makes the distinction between licensure as a function of the state to ensure that those practicing a profession have attained the minimum level of skills necessary to perform their functions safely and certification as a function of the profession to identify individuals within the profession who have attained particular skills or a level of proficiency above and beyond those minimally required. To a large extent, this distinction is lost within all aspects of the education enterprise. National Board certification (NBPTS, 2021) and opportunities for micro-credentialing (Berry, Airhart, and Byrd, 2016; French and Berry, 2017) are part of the solution, but a more formal process for supporting teachers' progression through the profession will also be critical to enhancing teaching literacy.

Circling Back to Related Knowledge and Skills

Even as I reframe assessment literacy in terms of the pedagogical content knowledge and the process of teaching, there remains a set of assessment- or measurement-related knowledge and skills that a competent teacher must possess. Research by DeLuca and his colleagues, for example, has identified specific assessment-related knowledge and skills that teachers appear lacking to function in a standards-based environment (DeLuca and Klinger, 2010; DeLuca and Braund, 2019). Xu and Brown (2016) would include this type of knowledge and skills as part of the knowledge base that forms the base of the TALiP pyramid . In DePascale et al (2018) we divided this set of knowledge and skills into three distinct bodies of knowledge: Testing or Test Design, Principles of Measurement (particularly an understanding of concepts related to uncertainly of individual student-level and aggregate scores), and selected areas of Data Analysis or Data Literacy.

From our perspective as educational measurement or assessment specialists we may have once viewed these knowledge and skills as the whole of assessment literacy. It should be clear that viewpoint was incorrect, or at least incomplete. Recent research on and conceptions of assessment literacy for teachers has demonstrated that, in practice, the core concepts and skills needed in each of the three areas listed above is relatively small and represents only a small portion of the varied knowledge and skills that teachers must possess to effectively use assessment to inform their decision making; that is, to display teaching literacy (Brookhart, 2011, DeLuca 2012; Xu and Brown, 2016).

As Popham (2018) stated, assessment literacy should focus on the "***small number of terribly significant understandings***", understandings that are "apt to influence educational decisions" and "understandings that a person has about the foundational procedures and concepts of educational

assessment." It is likely that many testing or test design skills that are needed by teachers can and should be taught in the context of other courses such as content-specific pedagogy, curriculum, or instructional methods (e.g., item types used for various purposes, content coverage, scoring student responses), and it would be preferable if that were the case. As with the knowledge and skills described in the previous section, instructional and training materials for teachers related to the knowledge and skills traditionally regarded as critical to assessment literacy can be designed to be framed within the context of teachers' decision making and other real-life tasks: for example,

- comparing changes in individual student performance from pre-test to post-test within a single unit or across units through the course of a trimester or year,
- evaluating the quality and performance of an individual item, task, or test through the use of qualitative and quantitative information.
- setting and evaluating acquisition of test-based goals and targets for student learning objectives,
- filtering a data set to examine the performance of a particular subgroup of students, or
- displaying student results graphically to aid in the interpretation of trends over time or comparisons between groups of students.

Other aspects of teaching closely related to assessment and assessment literacy such as grading and response to intervention have their own well-developed body of research, best practices, and instructional/training materials (McMillan and Nash, 2000; Brookhart, Guskey, Bowers, McMillan, Smith, and Welsh, 2016; Brookhart, 2017; Guskey and Brookhart, 2019).

Supporting Teachers' Interpretation and Use of External Large-Scale Tests

In this final section of the paper, we return to the topic of teachers' assessment literacy as it relates to their interpretation and use of results from large-scale tests. Given the disproportionate influence that large-scale tests have over curriculum, instruction, and local assessment, it is impossible to ignore them in a discussion of teaching literacy.

Large-scale tests provide information to inform very few teacher decisions, but those decisions are often very consequential. Decisions regarding student placement, remediation, and promotion or graduation all may be based solely on the results of external large-scale tests or are heavily influenced by student performance on those tests. In addition, the use of state tests for school accountability and educator accountability adds to the disproportionate influence of large-scale tests on teacher's decision making and actions in the classroom. For these reasons, it would seem obvious that assessment literacy for teachers must include the knowledge and skills necessary to appropriately interpret and use the results of external large-scale tests such as annual tests administered as part of the state assessment program, commercial interim assessments administered two to four times per year, and other commercial tests and inventories used as screening and placement tools.

However, with regard to teaching literacy (i.e., teachers' assessment literacy) the most significant "terribly significant understanding" that teachers must have about external large-scale tests is just how limited they are in providing information about individual student performance that can directly inform a teacher's instructional decisions. Such statements, often are interpreted as a criticism or weakness of large-scale tests, based on misconceptions about the purpose of the tests. Therefore, ensuring that teachers have an accurate understanding of the purpose of large-scale tests and the limits of information available from them must be the primary goal of any effort aimed at improving teachers' literacy with regard to large-scale tests. Beyond clarifying the purpose and use of large-scale tests, the best approach to supporting teachers' interpretation and use of large-scale tests may be through the enhancement of reporting of results from those tests.

Reporting Large-scale Test Results to Support Teaching Literacy

The primary responsibility for supporting teacher's appropriate interpretation and use of large-scale test results rests with the assessment vendors who provide the primary results from those tests. In the case of state assessment programs, the states also assume responsibility for conducting secondary analyses and providing context and resources to enhance teachers' ability to use the large-scale state test results appropriately. In both cases, the primary factor motivating reporting decisions should be how are these results going to be used to support teachers' decision making.

In the shift from norm-referenced to criterion-referenced testing we lost a sense of the fundamental purpose of reporting assessment results to teachers; that is, to convey information that will help those teachers interpret and use information from the test appropriately. For better or worse, norm-referenced "scores" were almost exclusively derived scores designed to provide useful information to teachers, administrators, and parents. In terms of the DIKW pyramid cited earlier in this paper, norm-referenced test scores might better be regarded as information (I) rather than data (D). Common examples of scores reported on norm-referenced tests include percentile ranks and stanines based on national and local norms, along with grade equivalent scores. Achievement levels or similar indicators of a desired level of mastery or proficiency were added to reporting in later years. Scale scores, which formed the basis for all of the reported scores across norm-referenced test forms, were rarely reported.

In contrast, scales scores, along with the achievement levels based on them, have become the primary metric used for reporting the results of criterion-referenced large-scale tests such as state tests. Scale scores, however, are clearly data rather than information. They have no inherent meaning or practical meaning and force teachers toward statistics-based interpretations which are of limited usefulness and which teachers, and most other intended users of large-scale test results, are ill-equipped to make.

It is incumbent on educational measurement and assessment specialists, therefore, to commit themselves, with the support of content experts and educators, to redesigning the reporting of results from large-scale tests so that they are more useful to teachers and other intended users. Prior to reporting results, assessment vendors, and states as appropriate, should also conduct all of the secondary analyses necessary to ensure that the test results are free of measurement bias and other factors that would have a negative impact on teachers' ability to use them. If it is not possible, to eliminate such factors, results should be reported in a way that accounts for their presence and supports teachers use of the results as reported.

In short, teachers should not be required to become statisticians or psychometricians to interpret and use results reported to them. Making the needed changes to reporting will do more to enhance teachers' literacy with regard to large-scale tests than perhaps any other action, with the possible exception of reducing the inordinate role that those tests play in K-12 public education.

As recommended in this paper for other aspects of assessment, results reported from large-scale tests should be designed to best support their intended use; that is, to support the decisions that must be made by the intended users. The logical outcome of such an approach is that the results reported to teachers will likely differ significantly from the results reported to district and school administrators, and from the results reported to state policymakers.

The outcome of a deliberative process focused on the how the results of large-scale test results may best be used by teachers to inform their decision making is likely to dramatically change the type of scores that are reported, the number of scores that are reported, and the manner in which scores are reported. It is beyond the scope of this paper to speculate as to what those redesigned results may be, but it would be surprising to see a scale score or standard error of measurement bar among them.

Acknowledgements

Thanks to Susan Brookhart and Carla Evans for their reviews of an earlier draft of this paper.

Cover image by Image by Anand Kumar from Pixabay

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Berry, B., Airhart, K. M., & Byrd, P. A. (2016). Microcredentials: Teacher learning transformed. *Phi Delta Kappan*, *98*(3), 34-40.

Bobbitt, F. (1918). Scientific method in curriculum-making. In Flinders, D. J., & Thornton, S. J. (2004). *The curriculum studies reader*. Psychology Press.

Brookhart, S.M. (2003). Developing measurement theory for classroom assessment purposes and uses. Educational Measurement: Issues and Practice, 22,4,5-12.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: issues and practice*, *30*(1), 3-12.

Brookhart, S. M. (2017). *How to use grading to improve learning*. ASCD.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., ... & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, *86*(4), 803-848.

Brookhart, S. M., & McMillan, J. H. (Eds.). (2019). *Classroom assessment and educational measurement*. Routledge.

Callahan, R.E. (1964). Education and the cult of efficiency. Chicago: University of Chicago Press.

Cochran, K. F., De Ruiter, J. A., & King, R. A. (1993). Pedagogical content knowing: An integrative model for teacher preparation. Journal of Teacher Education, 44, 263-272. doi:10.1177/0022487193044004004

Darling-Hammond, L. and Hyler, M.E. (2013) The Role of Performance Assessment in Developing Teaching as a Profession, in rethinking schools, 27,4. Downloaded from https://rethinkingschools.org/articles/the-role-of-performance-assessment-in-developing-teaching-as-a-profession/

DeLuca, C. (2012) Preparing Teachers for the Age of Accountability: Toward a Framework for Assessment Education, Action in Teacher Education, 34:5-6, 576-591, DOI: 10.1080/01626620.2012.730347

DeLuca C, Bellara A. (2013). The Current State of Assessment Education: Aligning Policy, Standards, and Teacher Education Curriculum. *Journal of Teacher Education*. 2013;64(4):356-372. doi:10.1177/0022487113488144

DeLuca, C., & Braund, H. (2019). Preparing Assessment Literate Teachers. In *Oxford Research Encyclopedia of Education*.

DeLuca, C. & Klinger, D.A. (2010) Assessment literacy development: identifying gaps in teacher candidates' learning, Assessment in Education: Principles, Policy & Practice, 17:4, 419-438, DOI: 10.1080/0969594X.2010.516643

DePascale, C. (1989). Factors influencing the cognitive emphasis of classroom tests. Paper presented at the 1989 annual meeting of the New England Educational Research Organization, Portsmouth, NH.

DePascale, C., Sharp, A., Ryan, K., and Betebenner, D. (2018). Building a Conceptual Framework for Assessment Literacy. Paper presented at the 2017 Annual Meeting of the National Council on Measurement in Education, San Antonio, Texas. Downloaded from https://www.nciea.org/node/517

Flood, J., Lapp, D., Squire, J. R., & Jensen, J. M. (2003). *Handbook of research on teaching the English language arts*. Lawrence Erlbaum Associates, Inc., Publishers, 10 Industrial Ave., Mahwah, NJ 07430-2262.

French, D. & Berry, B. (2017). Teachers, Micro-Credentials, and the Performance Assessment Movement. *Voices in Urban Education*, *46*, 37-43.

Frické, M. (2009). The knowledge pyramid: A critique of the DIKW hierarchy. Journal of Information Science,

Greenberg, J., McKee, A., & Walsh, K. (2013). "Teacher prep review: A review of the nation's teacher preparation programs." Available at SSRN 2353894.

Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. ***Educational Measurement: Issues and Practice, 4***(1), 15–18. https://doi.org/10.1111/j.1745-3992.1985.tb00286.x

Guskey, T. R., & Brookhart, S. M. (2019). *What we know about grading: What works, what doesn't, and what's next*. ASCD.

Henry, N.L. (1974). Knowledge management: A new concern for public administration. Public Administration Review, 34,3,189-196.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do. Phi Delta Kappan. 89,2,140-145.

Heritage, M. and Heritage, J. (2013). Teacher questioning: The epicenter of instruction and assessment. Appied Measurement in Education. 26,3,176-190.

Heritage, M. and Vendlinski, T. (2006). Measuring teachers' mathematical knowledge. CSE Technical Report

696. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE), Graduate School of Education & Information Studies. University of California, Los Angeles.

Heritage, M., Kim, J., Vendlinski, T.P., and Herman, J.I. From evidence to action: A seamless process in formative assessment. (2008) CRESST Report 741. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE), Graduate School of Education & Information Studies. University of California, Los Angeles.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking Pedagogical Content Knowledge: Conceptualizing and Measuring Teachers' Topic-Specific Knowledge of Students, *Journal for Research in Mathematics Education JRME*, *39*(4), 372-400. Retrieved Mar 2, 2021, from https://pubs.nctm.org/view/journals/jrme/39/4/article-p372.xml

Hill, H.C., Schilling, S.G. and Ball, D.L. (2004). Developing measures of teachers' mathematics knowledge for teaching. The Elementary School Journal. 105,1,11-30.

Joint Committee on Standards for Educational Evaluation (2015). Classroom Assessment Standards for PreK-12 teachers.

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. ***Evaluation & the Health Professions, 17***(2), 133–159. https://doi.org/10.1177/016327879401700202

Kane, M. T., & Wools, S. (2019). Perspectives on the Validity of Classroom Assessments. *Classroom Assessment and Educational Measurement*, 11.

Kliebard, H. M. (2004). *The struggle for the American curriculum, 1893-1958*. Psychology Press.

Magnusson S., Krajcik J., Borko H. (1999) Nature, Sources, and Development of Pedagogical Content Knowledge for Science Teaching. In: Gess-Newsome J., Lederman N.G. (eds) Examining Pedagogical Content Knowledge. Science & Technology Education Library, vol 6. Springer, Dordrecht. https://doi.org/10.1007/0-306-47217-1_4

Mann, H. (1840). On the art of teaching. 1989 edition. Applewood Books.

McMillan, J. H., & Nash, S. (2000). Teacher Classroom Assessment and Grading Practices Decision Making.

McMillan, J.H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. Educational Measurement: Issues and Practice, 22, 4, 34-43.

Messick, S. (1990). Validity of test interpretation and use. ETS Research Report Series. 1990,1, 1487-1495.

Moss, P.A. (1994). Can there be validity without reliability? Educational Researcher, 23,2,5-12.

Moss, P.A. (2003). Reconceptualizing validity for classroom assessment. Educational Measurement: Issues and Practice, 22,4,13-25.

National Board for Professional Teaching Standards (2021). National Board Certification. Downloaded from https://www.nbpts.org/national-board-certification/

Penuel, W. R., & Shepard, L. A. (2017). Social models of learning and assessment. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 146-173.

Popham, W.J. (2011). Assessment Literacy Overlooked: A Teacher Educator's Confession, The Teacher Educator, 46:4, 265-273, DOI: 10.1080/08878730.2011.605048

Popham, J. (2018). What is assessment literacy? Michigan Assessment Consortium. Learning Moment. Downloaded from https://www.michiganassessmentconsortium.org/event/assessment-learning-network-2018-19-event-1/

Shepard, L. A. (2019). Classroom assessment to support teaching and learning. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 183-200.

Shepard, L. A. (2019). Should "Measurement" Have a Role in Teacher Learning about Classroom Assessment?. *Classroom Assessment and Educational Measurement*, 192.

Shepard, L. A., Penuel, W. R., & Davidson, K. L. (2017). Design principles for new systems of assessment. *Phi Delta Kappan*, *98*(6), 47-52.

Shulman, L.S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15,2, 4-14.

Shulman, L.S. (1987). Assessment for teaching: An initiative for the profession. The Phi Delta Kappan. 6*9,1,38-44.*

Shulman, L.S. (1998). Theory, practice, and the education of professionals. The Elementary School Journal. 98,5,511-526.

Shulman, L.S. and Shulman, J.H. (2008,2009). How and what teachers learn: A shifting perspective. The Journal of Education. 189,1/2,1-8.

Gudmundsdottir, S. and Shulman, L. (1987) Pedagogical Content Knowledge in Social Studies, Scandinavian Journal of Educational Research, 31:2, 59-70, DOI: 10.1080/0031383870310201

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational measurement: Issues and practice*, *22*(4), 26-33.

Traub, R.E. and Rowley, G.L (1991). Understanding reliability: An NCME instructional module. Educational Measurement: Issues and Practice, 10, 1, 37-45.

Wylie, C., & Lyon, C. (2016). Using the formative assessment rubrics, reflection and observation tools to support professional reflection on practice (Revised). *Formative Assessment for Students and Teachers*.

Xu, Y., & Brown, G. T. L. (2016). Teacher Assessment Literacy in Practice: A Reconceptualization. Teaching and Teacher Education, 58, 149-162.