Formative Assessment: Literature Review

Tonja Jarrell, MA, PhD Candidate

Graduate School of Education

Policy, Organization, Measurement, and Evaluation

University of California, Berkeley

Contents

Abstract

The purpose of this literature review is to uncover how formative assessment can be used to understand student progress and improve instruction in mathematics. The work finds that a lack of policy support contributes to classroom teachers' unfamiliarity with formative assessment and their varying levels of understanding of the purpose of the assessment. The paper begins by clarifying terms to advance the collective understanding of the language of assessment in order to promote a deeper understanding of the purpose of formative assessment: to *uncover* how students reason in mathematics. By refining a community definition of formative assessment, this review argues that student achievement can be increased through clarity of purpose. Understanding the purpose of assessment allows the teacher to create transparency for students about how and why the evidence will be used in their learning and helps the teacher focus on student learning rather than on normative achievement. The method for the review does not focus on developing a typology of assessment but broadly examines two bodies of research by beginning with a clarification of assessment terms, then exploring assessments to support learning, and finally applying the implications outlined by each researcher to future research, human capital development, and product development.

This review merges from two separate research bodies—assessment and student reasoning and assessment in mathematics—the purpose of defining the language of assessment and formative assessment, and the ontological shift from using assessment (e.g. summative) to measure student outcomes to using assessment (e.g. formative and diagnostic) to support the teaching and learning of mathematical reasoning. The review

finds the differences between formative, summative, and evaluative are determined based on the uses of the assessment findings and concludes that the same data may be reviewed for different purposes. The synthesis of research concludes that formative assessment has a single clear purpose: to help learning and teaching. The research suggests that an assessment is formative only when the evidence is used to decide where the students are in their learning, where they need to progress to, and what the best strategies are to get them there. This review argues that what matters most in mathematical assessment is what the teacher learns about how the student reasons and his or her numerical proficiency, not whether the assessment itself is labeled as formative or summative. It is the goal of the latter portion of the literature review to deepen understanding of why and how instructional tasks and alternative forms of formative assessment that include discourse can be an integral part of increasing students' reasoning and sense making in mathematics. However, the paper finds that this characterization of instructional environments stands in sharp contrast to the ways most classrooms are organized and the methods by which most teachers assess student reasoning. These findings hold important implications for policy makers as well as practitioners in regard to pre-service training, professional development, and product development supporting change in the field.

Introduction and Background

There is an analogy so often and widely cited in the education community that it has taken on the status of a near urban legend: "When the cook tastes the soup, it's formative evaluation, and when the guest tastes the soup, it's summative evaluation" (Joyner, 2011).

While this quote can be attributed to Robert Stake in his 2004 publication *Standards-Based and Responsive Evaluation*, the many who employ the quote often mistake the analogy as being sufficient to bring clarity to the terms contained within.

The need for clarity of language around assessment is especially critical today, as the current policy environment of high-stakes testing has thrust assessment, primarily in mathematics and English language arts, to the forefront of education. The adoption of the No Child Left Behind Act of 2002 gave rise to a focus on accountability through the testing of children's mastery of content standards from elementary through high school in mathematics and English. This focus rapidly shifted the institutional and cultural logic around tests and spawned a new vernacular to capture the range of formative and summative testing. Why, when, and how we assess students have an impact on their educational experience and their learning (Harlen, 2007b). The terms and purposes of assessment types may vary, but what unites the conversation around assessment is the shared aim to understand student progress and improve instruction for all students.

Researchers in the United States and abroad have found that high-stakes testing now dominates the dialogue in the field of assessment. As Black and Wiliam (1998) found, the emphasis is on providing overall summaries of normative achievement rather than on using data to make a strategic diagnosis about students' mathematical understanding and future needs. As a result, these initiatives leave teachers with inadequate models for formative assessment. In the United States this reform movement has been characterized by a distrust of teachers and an accepted cultural norm that external, summative testing will, on its own, improve learning results (Black, 1998). This lack of policy support for high-quality formative assessments is an integral factor that contributes to a gap in supporting

the development of alternative forms of assessment and teacher training aimed at measuring students' reasoning in mathematics. This policy climate leads to teachers who may then be unfamiliar with the purposes of and uses for formative assessment and professional communities of teachers that have varying, often conflicting, understandings of assessment.

In *Assessment Standards for School Mathematics*, the National Council of Teachers of Mathematics states that the main goal of assessment is "to advance students' learning and inform teachers as they make instructional decisions" (Mathematics, 1995). Additionally, the Common Core State Standards (CCSS) for Mathematical Practice call designers of curricula, assessments, and professional development to create measurements of student reasoning to evaluate students' mathematical proficiency (NGA Center and CCSSO 2010). The authors of the CCSS for Mathematics make the case that teaching and assessing reasoning are the very essence of mathematics instruction and are the benchmark on which student learning is centered. It is this call for instruction and assessment designed to evaluate students' reasoning that permeates the standards for mathematical content and the empirical papers reviewed within.

It is, then, the aim of this conceptual literature review to clarify the terms and purpose of formative assessment in order to bring understanding and clarity to the field of mathematics instruction. The secondary goal is to understand how assessment can be designed to uncover students' reasoning in mathematics in order to improve instruction. These two bodies have not previously been conversant; thus it is the work of this review to allow these two bodies to speak to each other to inform the work of policy makers, developers, and practitioners.

As Black and Wiliam state about the confusing language surrounding formative assessment, "There is no internationally agreed upon term here" (Black, 1998). The intended outcome of this review may guide policy makers and educators in changing both policy and classroom practice to use formative assessment data to support, or inform, instructional decisions based on students' sense making on assessments.

Finally, Harlen's work in "Criteria for Evaluating Systems for Student Assessment" (2007a) shows us that an effective assessment system is an open one, where all involved know what evidence is used and what judgments are to be made. Much of the emotion aroused by assessment in schools and classrooms is a result of fear or suspicion as to the purpose and utility of the assessment. To eliminate this wariness of assessment, the education community must be completely transparent about the need for and purpose of assessment and why it is carried out in particular ways. Even the youngest of students can be given a developmentally appropriate explanation of what evidence they and their teachers will use to judge the progress they are making. This communication of purpose helps students take part in assessing their own work, which is a key feature of using assessment to improve learning (Harlen, 2007b).

Problem Statement: Clarifying Terminology and Shifting Focus

While there seems to be a pervasive, unspoken agreement in education that we are all speaking the same language of assessment, in reality, there is a lack of shared understanding of common terms, and therefore an ill-defined purpose. This may lead to

misuse of assessments and inaccurate conclusions drawn from the data, which ultimately impacts a teacher's ability to effectively improve student learning.

In Harlen's (Harlen, 2007a) empirical work on evaluating systems for student assessment, the author posits that negative impacts can occur from a formative assessment when teachers follow procedures mechanically without a deep or thorough understanding of the assessment's meaning or purpose. This finding is significant to the purposes of this review because it creates a necessary sense of urgency for reviewing the extant body to develop a shared understanding of formative assessment. Harlen's work illustrates that a shared definition of formative assessment is larger than simple semantics; a shared language is necessary to understand the purpose of the assessment as well as ensure the rigorous and thoughtful implementation of each assessment. Without a shared language, teachers and students may not fully utilize feedback to improve instruction and increase learning.

Harlen further explains that there are two main reasons for assessing students: to help their learning and to report on what they have learned. He argues that researchers typically discuss these reasons as different purposes for assessment and "mistakenly as different *kinds* of assessments that are somehow opposed to one another" (Harlen, 2007b). The notion that these two goals of assessment are different results in teachers not seeing the connection between how students learn and how students demonstrate what they have learned. Therefore, another goal of this review is to assist the education community in moving from the idea of assessment *of* learning to the idea of assessment *for* learning. That is, the review aims to help educators use formative assessment strategically to improve instruction for all students.

To illustrate this principle in the practitioner literature, Joyner and Muri (2011) find that there are six key questions teachers must routinely ask during any assessment. Three of the questions—Why are we assessing? When do we assess? and How are we assessing?—cannot be clearly answered without a deep, working understanding of the language of assessment. Can a teacher know when to use an online interview as a formative assessment of students' mathematical reasoning if he or she does not thoroughly comprehend the differences in purpose between formative and summative assessments? This review argues that this teacher would be unable to utilize such an assessment effectively and therefore would be unable to make decisions about the resulting data that would improve student learning.

Across the existing body of research on mathematics assessment, terms of assessment are used in differing, often confusing ways that serve to obfuscate the purpose of the assessments themselves. For example, one source uses terminology in the following manner: "our approach includes . . . summative assessments and interim assessments as well as formative resources to monitor student progress" ("SMARTER Balanced Assessment Consortium," 2012). From such writing it becomes apparent that the main purpose of this review, to move the collective understanding forward, is to create clarity of terms around the language of assessment. This is necessary to promote a deeper understanding of the purpose of formative assessment: to *uncover* how students reason in mathematics. As Harlen (2007a) explains, the type of assessment itself is not the distinguishing characteristic but, rather, the purpose of the evidence gathered from the assessment.

Perhaps more important than the distinctions between summative and formative are the nuances in the types of formative assessments for the data they can give teachers to make decisions about instructional practices and learning. While this review attempts to identify the differences between summative and formative assessment, its primary goal is to move the research body forward by examining how mathematics assessments can help practitioners understand how students understand and reason. As the developers of the CCSS for Mathematics explain, assessments are a method to evaluate student understanding at the intersection of mathematical practice and content (CCSSM, 2011).

While assessment in mathematics has been pushed to the forefront by accountability policies, it is certainly not a new topic. We can look back through attempts to initiate formative assessment, for example, over the years and trace changes in the nomenclature (Joyner, 2011). Formative assessment has been called *continuous assessment*, so how did the term *formative* gain traction and how can we come to agreement as a professional community on a shared understanding of the meaning? Going back to the Latin root of *assessment*, we learn that the term *assessus* means "to sit beside." In present-day English, Merriam-Webster's Collegiate Dictionary lists one definition of *formative* as "capable of alteration by growth and development." (Merriam-Webster, 1997)This definition proves crucial in helping to meet the other goal of this review, which is to shift our thoughts from using formative assessment to assess learning to using it to *support learning*. By refining a community definition of formative assessment, this review argues that student achievement can be increased through clarity of purpose. Understanding the purpose of the assessment allows the teacher to create transparency for students about how and why the evidence will be used in their learning and helps the teacher focus on

student learning rather than normative achievement (Harlen, 2007b). Wiliam's research supports this argument. He writes, "the use of day-to-day formative assessment is one of the most powerful ways of improving learning in the mathematics classroom" (Wiliam, 2004).

Design and Methods of Review

The purpose of this analytic literature review necessitated a search of both empirical and practitioner research in the extant bodies of assessment and assessing reasoning and sense making in mathematics. With such a wide net cast, it is difficult to approach the review with any a priori theories. This review does not focus on examining the methods of the studies nor the theoretical foundations of the studies reviewed herein but, rather, synthesizes the concepts and findings in the research in an effort to understand the implications for both researchers and practitioners in the field of assessment through agendas built for teacher professional development and product development. The method for this conceptual review does not focus on developing a typology of assessment but on broadly reviewing the two bodies by beginning with a clarification of assessment terms, exploring assessments to support learning, and applying the implications outlined by each researcher to future research, human capital development, and product development.

The Organizing Principle: Framework and Related Literature

This review merges from two separate research bodies—assessment and student reasoning and assessment in mathematics—the purpose of defining the language of

assessment and formative assessment, and the ontological shift from using assessment (e.g. summative) to measure student outcomes to using assessment (e.g. formative and diagnostic) to support the teaching and learning of mathematical reasoning. Merging these two bodies may prove to have significant implications on future research agendas as well as teacher training and product development. The review seeks to answer the following empirical questions:

- What is the meaning of the term *formative assessment*?
- What is the purpose of formative assessment?
- Can the use of formative assessment improve instruction? Under what measures can teachers use assessment to understand students' mathematical reasoning to support learning?

Assessment Research: Contributions and Concerns

Similar to the National Council of Teachers of Mathematics' stated goal for assessment as being able to inform students and teachers of progress, it is the primary goal of this review's analysis to inform the field of the variances in the language used to define formative assessment tools. Studies have been synthesized to identify the variations across terms and to develop a shared definition of formative assessment. A secondary goal of this review is to answer whether formative assessment can improve instruction and, if so, under what measures do practitioners shift gears from using formative assessments to

measure results to using formative assessments to improve students' achievement (Wiliam, 2004).

Initial analysis reveals that much of the assessment research falls into two neat groups: (1) anecdotal research that directs practitioners on how to implement a formative or summative assessment most effectively or efficiently and (2) empirical research that tests construct validity on the type of assessment implemented. Joyner and Muri's (2011) work is an example of the former while Rea-Dickins and Gardner's (2000) paper is an example of the latter. However, this review seeks to pull out from the assessment research the findings of the studies reviewed to contribute to the body a deeper understanding and greater clarity of the language and purposes of formative assessment.

Rea-Dickins and Gardner (2000) open their empirical case study with the broad purpose of assessment stated as a function that confirms the appropriateness of teaching. Perhaps, then, Rea-Dickins and Gardner also offer a more nuanced definition of formative assessment than other studies when they argue that the differences between formative, summative, and evaluative are determined based on the uses of the assessment findings and conclude that the same data may be reviewed for different purposes. Further, the authors find that formative assessments are iterative in nature and are used as tools to tune instruction to be responsive to students' learning needs, as evidence of learner attainment matched against national curriculum targets and tests, and as evidence for evaluation of teaching. Rea-Dickins and Gardner's research offers a more complex account of the defining features of formative and summative assessments than is provided in much of the extant body on assessment (Genesee, 1996).

Harlen's empirical paper, while focused chiefly on developing criteria to evaluate the advantages and disadvantages of particular assessment procedures, defines assessment as "the process of deciding, collecting, and making judgments about evidence relating to students' achievement of particular goals of learning" (Harlen, 2007a). That is to say, assessment is about student achievement and the relation to individual or common learning goals. Harlen's work stands out in this review, as it is the only paper to identify the exact time when the author believes assessment became distinguished into four different purposes: formative, diagnostic, summative, and evaluative. Harlen finds that this typology became widespread in 1988. He further explains that *formative* was the term used to identify assessments that provide evidence of student achievement in relation to their learning goals. He distinguishes summative assessment by stating that it is a summary of achievements at a particular time. Perhaps the most significant contribution from Harlen for this review is his assertion that formative assessment is "essentially a pedagogical approach" (2007a, rather than a separate activity outside of teaching. He states that formative assessment has a "single clear purpose" (16): that of helping learning and teaching. The implication of that approach in practice looks like evidence being gathered during learning activities and being interpreted relative to student progress toward the lesson goals. The author argues that if the assessment does not serve this purpose, it is not, by definition, formative. Finally, Harlen concludes that the various types and purposes of assessment are not independent of one another but are all part of an interconnected system used to make decisions about students, with each part able to exert influence over how each individual assessment is used and interpreted.

The interpretation and use of assessment data is the central focus of Black and Wiliam's meta-analysis of 250 studies of assessment. In this ten-year project, the authors found that teachers who focused on assessment for learning, as opposed to the traditional assessment of learning, produced a substantial increase in their students' achievement (Black, 1998; Wiliam, 2004). What Black and Wiliam found, however, that sets their work apart from the other papers reviewed is their typology of the key ingredients of formative assessment. These ingredients are effective questioning, feedback, an understanding of the criteria for success on the part of the learners, and peer- and self-assessment. Wiliam (n.d.) then pushes this thinking forward to outline the idea of "regulation of learning." Wiliam's findings hold important implications for pre-service training, professional development, and product development. He writes about the key ingredients of formative assessment, "Feedback is not the same as formative assessment. . . . [F]eedback is formative only if the information fed back to the learner is used by the learner in improving performance. . . . [I]f the information fed back to the learner cannot be used by the learner in improving their performance it is not formative" (8).

Black and Wiliam's seminal meta-analysis of 250 sources from the research literature creates a strong case for policy makers and practitioners alike to question their goal of improving student performance. For, as the authors maintain, formative assessment is not a magic bullet for education. It is a complex process that requires extensive professional development but one that can raise standards for all students (Black, 1998). They broadly define an assessment as being formative when the evidence is used to adapt the teaching to meet student needs. More importantly, their analysis supports the finding that strengthening the practice of formative assessment produces significant and

substantial learning gains. Their most important conclusion, however, is that improved formative assessment helps low achievers more than other students so that achievement raises overall in addition to closing the opportunity gap. Closing this gap through evolved formative assessment practices is not simple. Black and Wiliam maintain that the largest difficulties with formative assessment revolve around three central issues: poorly executed testing practices such as questioning encourage rote and superficial learning; feedback is not utilized to increase personal improvement; and formative feedback serves a primarily managerial function in most classrooms.

All authors agree that an assessment is formative only when the evidence is used to decide where the students are in their learning, where they need to progress to, and what the best strategies are to get them there. Decisions about what data to collect, how the data should be used, and by whom follow from the reasons for the assessment, not the type of assessment itself. In practical terms, this review concludes that what matters most in mathematical assessment is what the teacher learns about how the student reasons and his or her numerical proficiency, not whether the assessment itself is labeled as formative or summative.

Table 1: Characteristics of Assessment Found in the Literature

Author(s)	Purpose of Assessment	Definition of Formative Assessment	Definition of Summative Assessment	Conceptual Dimensions of Terms Used	Additional Assessment Terms Defined
Joyner and Muri (2011)	• Determine what students know and can	• Capable of alteration by growth and	• None	• Latin definition of *assessus* is "to	• Diagnostic • Pre-test

Author(s)	Purpose of Assessment	Definition of Formative Assessment	Definition of Summative Assessment	Conceptual Dimensions of Terms Used	Additional Assessment Terms Defined
	do, advance student learning • Identify students' understandings and misunderstandings	development		sit beside"	
Rea-Dickins and Gardner (2000)	• Confirm the appropriateness of teaching • Provide data for planning of appropriate language support • Play a role in managing, monitoring, and promoting student learning	• Iterative in nature • Used to tune instruction to be responsive to students' learning needs • Evidence of curricular learning and development • Evidence of learner attainment matched	• Required for administrative purposes • Used to assign grades • Measures attainment and proficiency	• Formative as contrast to summative • Focus on increasing validity and reliability of classroom formative assessments • The type of assessment is determined by how the findings are used	• Evaluative: assessment where data is used for normative purposes

Author(s)	Purpose of Assessment	Definition of Formative Assessment	Definition of Summative Assessment	Conceptual Dimensions of Terms Used	Additional Assessment Terms Defined
		against national curriculum targets and tests • Evidence for evaluation of teaching			
Genesee and Upshur (1996)	• Inform instructional decision making	• Continuous assessment • Able to be used informally by classroom teachers • Often occurs on daily basis	• None	• Focus on formative evaluation for language instruction	• None
Harlen (2007a, 2007b)	• Dependent upon how assessment is carried out • Pedagogical approach • Make judgments	• Helps learning and teaching • Provides evidence of student achievement in relation to students'	• A summary of achievements at a particular time • Used for keeping records and giving	• Focus on construct validity • Finds that reliability criteria are more readily met with	• Evaluation: the process used for making decisions and judgments about systems,

Author(s)	Purpose of Assessment	Definition of Formative Assessment	Definition of Summative Assessment	Conceptual Dimensions of Terms Used	Additional Assessment Terms Defined
	about evidence relating to students' achievement of learning goals	learning goals	progress reports to various stakeholders	teacher assessments than with external tests	programs, and processes • Diagnostic • Evaluative
Wiliam (n.d., 2007)	• Improve student performance	• Includes key ingredients: effective questioning, feedback, an understandin g of the criteria for success on the part of learners, peer- and self-assessment • Feedback must contain a recipe for future action in order to be formative	• None	• Regulation of learning • Formative lessons are those that change in light of evidence about student progress	• Feedback
Black and Wiliam	• Provide	• Assessment	• None	• Unique meta-	• Poverty of

Author(s)	Purpose of Assessment	Definition of Formative Assessment	Definition of Summative Assessment	Conceptual Dimensions of Terms Used	Additional Assessment Terms Defined
(1998)	information to be used as feedback to modify teaching and learning	whose evidence is used to adapt the teaching to meet student needs		analysis that focuses on changing policy environment	practice: formative assessment practices are beset with problems and shortcomings

Conceptual Dimensions of Assessment and Technology Research

Published empirical research on the use of technology in assessment saw a dramatic increase when the Common Core State Standards (CCSS) were initiated in 2009, as one of the chief challenges in implementing the CCSS is developing "next generation" assessment systems to evaluate the higher-order learning called for in the CCSS (Levin, 2011). The additional surge in studies of online assessments occurred when the United States Department of Education provided over $350 million to fund four national assessment consortia (Levin, 2011). Publications from these consortia, such as Smarter Balanced and PARCC (Partnership for Assessment of Readiness for College and Careers), share research from working groups committed to using technology to develop assessments aligned to the CCSS. The most robust commonalities across the nationwide consortia for collaboratively developing computer adaptive testing (CAT) are three rationales: increased efficiency and

security, more detailed information for teachers, and advances in accountability through accurate evaluations of student progress and achievement ((PARCC), 2012; SMARTER Balanced Assessment Consortium," 2012).

Assessment Practices That Expose Mathematical Reasoning

In her 2003 book, Lampert found that mathematics is associated with certainty: knowing it and being able to get the right answer quickly (Lampert, 2003). Through their rules, instruction, and assessment routines, teachers shape these cultural experiences for students. It is the goal of this portion of the literature review to uncover the instructional and assessment practices that reveal students' mathematical reasoning. This process is inclusive of problem-solving and sense-making strategies such that the evidence gained may be used to meet students' learning needs.

In the 1980s, the National Council of Teachers of Mathematics originally heralded in a policy agenda of problem solving that impacts teacher education and teachers' classroom focus today; "problem solving must be the focus of mathematics" for K–12 mathematics (Schoenfeld, 1992). Schoenfeld argues that it is logical to conclude that the primary goal of mathematics instruction should be to have students become highly competent problem solvers. Yet his work shows that this goal is unclear as there have been multiple interpretations of just what problem solving is and how to assess it. Schoenfeld offers a helpful analogy of mathematics instruction that has been adopted as the view of this literature review: mathematics instruction and assessment without a focus on problem solving and sense making is akin to instruction in English that focuses exclusively on

grammar and ignores reading comprehension. The field has since moved forward to include the NGA Center and CCSSO's (2010) call for mathematical standards that are informed by the most effective models and provide educators with a common understanding of what appropriate benchmarks are for teaching and assessing reasoning as the foundation of mathematical proficiency.

It is the goal of this latter portion of the literature review to deepen understanding of why and how instructional tasks and alternative forms of formative assessment that include discourse can be an integral part of increasing students' reasoning and sense making in mathematics. This review adopts Boesen, Lithner, and Palm's definition of reasoning, "the line of thought adopted to produce assertions and reach conclusions" (Boesen, 2010), as well as their definition of assessment task, "an instruction of question that requires a student response under certain conditions and specific scoring rules" (92).

Schoenfeld (1992) argues that, from this perspective, learning mathematics is empowering for students, and he finds that mathematically powerful students are those who are quantitatively literate. Schoenfeld concludes that in addition to using problem-solving strategies, engaging in mathematical practices, such as explaining to themselves or others why a mathematical rule is true or where a mathematical rule comes from, is a fundamental aspect of thinking mathematically. Schoenfeld finds that in order for a teacher to understand a student's behavior when assessing his or her mathematical reasoning ("which options are pursued—in which ways" [339), the teacher needs an assessment task to reveal not only what information the student possesses but also how the student accesses that information and uses it. He goes on to explain that in assessing a student's mathematical decision making and reasoning, the teacher needs to know what options the

student had available. That is, did a student fail to pursue particular options because he or she overlooked using them, or was the student completely unaware those solutions existed? He argues that this type of assessment of students' reasoning is important for understanding their misconceptions and misunderstandings. A teacher would, hopefully, make different instructional decisions about the future needs of students based on their answers to the previous question. This can be problematic because the dominant routine of instruction in math classes is characterized by teachers explaining new material, solving problems on the board or overhead, and assigning students to work on problems on their own, or what Burkhardt calls the "exposition, examples, and exercises mode" (Burkhardt, 1988). This mode is true even for instruction to increase reasoning and problem solving, as the teaching of mathematical problem solving is as much work for the teacher as it is for the students. Schoenfeld illustrates this as he describes the three questions he asks students to assess their reasoning (1992, 336):

1. What exactly are you doing (Can you describe it precisely)?
2. Why are you doing it (How does it fit into the solution)?
3. How does it help you (What will you do with the outcome when you obtain it)?

Schoenfeld's illustrative point demonstrates that by virtue of these oral assessment questions, students *gave themselves the opportunity* to solve the problem. Schoenfeld summarizes his research by calling for behavior modification of students to unlearn inappropriate mathematical behaviors from prior instruction. He concludes that the task of modeling and creating the right instructional and assessment context is challenging for

classroom teachers and will require substantial professional development and pedagogical coaching. Finally, Schoenfeld's work reminds us that "a teacher's sense of the mathematical enterprise determines the nature of the classroom environment . . . that, in turn, shapes students' beliefs about the nature of mathematics" (1992, 368). This explains why often teachers in this era of high-stakes accountability profess a belief in discovering where there students are in relation to mathematical reasoning prior to instruction but, under the pressure of content coverage, sacrifice this goal for rote drilling and multiple-choice tests.

Boesen, Lithner, and Palm (2010) also examine student learning by studying the relationship between types of assessment tasks and students' mathematical reasoning. Their results show that when assessment tasks are related to textbook tasks, students attempt to solve the problems by trying to recall facts or algorithms (a practice called imitative reasoning by the authors) because the tasks do not require conceptual understanding. The authors explain imitative reasoning as memorized reasoning, recall, and recalling an algorithm or sequence of rules. Conversely, when the assessment tasks do not resemble textbook tasks, the assessment items elicit what the authors call creative mathematically founded reasoning from the students. This study showed that most teacher-made tests focused heavily on algorithmic procedures and did not provide extensive opportunities for displaying different kinds of mathematical reasoning. Further, the authors did not find assessment of sense-making strategies that allow students to become problem solvers. This is problematic, the authors argue, because assessments are a cornerstone in students' work and "influence students by directing their attention to particular aspects of content and specifying ways of processing information" (103). This finding ties to Wiliam's (2007) conclusion that assessments, especially those used

formatively, can affect students' learning through the assessment tasks' influence on students' reasoning. In formative assessments, the kind of reasoning that the questions elicit is important for student learning because this is the very reasoning that the feedback can address (Wiliam, 2007). Boesen and his colleagues' research may be helpful in both the development of an assessment task intended to measure student reasoning and the analysis of assessment tasks and the interpretation of their results. An extension of this for practitioner use would be to understand the reasons why some students try to solve certain assessment tasks with imitative reasoning rather than creative mathematically founded reasoning. Perhaps the most important finding from Boesen et al.'s work for the purposes of this literature review is that mathematical assessments that use only written responses can miss important aspects of students' reasoning, such as misunderstandings and misconceptions. Further, for developers and teachers alike, it is helpful to know that teacher-made tests display a heavy emphasis on imitative reasoning and this may have to do with teacher beliefs. In Boesen and his colleagues' study, teachers stated the main reason for the focus on test items that elicit imitative reasoning is that they do not believe that low-performing students are able to learn and use creative mathematically founded reasoning. These teachers stated that their tests reflect their teaching, leaving us with the conclusion that in such classrooms low-performing students and the rest of their classmates receive a mathematical education that is focused on imitative reasoning.

Burton's (1984) work supports Schoenfeld's assertions that mathematical thinking is not thinking about the subject matter itself but a style of thinking that is a function of particular operations, processes, and dynamics. Burton's premise is that because classroom instruction and assessment become confused with thinking about mathematics and

mathematical thinking, there has been a failure in separating mathematical content from mathematical practice for students. Burton's empirical question is, Can mathematical thinking be taught and measured? His findings state that mathematical thinking, reasoning, and sense making can be taught if teachers are trained and practice spending increased time on process and decreased time on content. Stein and Lane discuss Burton's ideas as the use of instructional tasks and assessments that "engage students in the doing of mathematics" (M. K. Stein, & Lane, S., 1996). Their research shows that student performance gains were seen in classrooms where instruction and assessment tasks implemented the use of multiple solution strategies, multiple representations, and explanations. Further, declines were seen in classrooms where these tasks did not occur and where students had little or no mathematical communication. Their work presents compelling arguments for instructional environments that are characterized by an emphasis on mathematical thinking, reasoning, and problem solving. Their work links directly back to Schoenfeld's call for instructional tools and assessments that support teachers in fulfilling mathematical reasoning as their goal rather than mathematical content coverage.

In her more recent work with Henningsen, Stein researches classroom environments that develop students' capacities to "do mathematics," to engage actively in high-level mathematical thinking and reasoning (Henningsen, 1997). Their findings suggest that there are specific factors present in classroom tasks and assessments when students' engagement is successfully maintained at high levels. They build on Schoenfeld's earlier writing, where he describes a dynamic and exploratory stance toward the discipline of mathematics. This stance requires teachers and students to focus on the active, generative

processes engaged in by doers and users of mathematics rather than to view mathematics as a static system of facts and concepts (Henningsen, 1997). Their work supports and builds on these ideas by arguing that such active mathematical processes involve the use of specific mathematical reasoning tools to explore patterns, frame problems, and justify processes. This has implications for classroom teachers and their ideas about what students need to learn and how to assess them. If classrooms are to develop these capacities in students, they must provide frequent opportunities to engage in worthwhile mathematical and assessment tasks that include looking for and exploring patterns to understand mathematical structures and underlying relationships, using available resources to formulate and solve problems, thinking and reasoning in flexible ways, and justifying and communicating one's mathematical ideas. Henningsen and Stein argue that the nature of classroom tasks and assessments can influence and structure the way students think and can broaden or limit their views of the subject. The authors supply a specific example in their work: "Teachers should be especially attentive to the extent to which meaning is emphasized and the extent to which students are explicitly expected to demonstrate understanding of the mathematics underlying the activities in which they are engaged. . . . [E]xplicit connections between the mathematical ideas and the activities in which they are engaged must be frequently drawn and assessed" (Henningsen, 1997).

Henningsen and Stein come together in a paper with Grover to focus on the mathematical tasks identified as important vehicles for building student capacity for mathematical thinking and reasoning (M. K. Stein, Grover, B.W., & Henningsen, M., 1996). Their empirical question began as, What types of instructional environments might reasonably be expected to produce these kinds of student outcomes? Their analysis

examined a stratified random sample of 144 mathematical tasks in terms of task features (the number of solution strategies, number and kind of representations, and communication requirements) and cognitive demands (memorization, use of procedures, "doing of mathematics") and found that while teachers were setting up task features consistently, they tended to allow the cognitive demands of high-level mathematical tasks to decline in assessments. This is similar to national normed tests where students' mathematical reasoning is not assessed because it is not easily quantifiable in a fiscally expeditious manner. The authors recommend that classrooms expose students to meaningful and worthwhile mathematical tasks that are truly problematic for students rather than "simply a disguised way to have them practice an already demonstrated algorithm" (456). They conclude that such tasks are characterized by features such as having more than one solution strategy, being able to be represented in multiple ways, and demanding that students communicate and justify their procedures and understandings in both written and oral form. The findings suggest requiring oral explanations because correct answers can sometimes hide confusion and misconceptions, and equally as often, incorrect responses can hide what students actually do understand (M. K. Stein, Grover, B.W., & Henningsen, M., 1996). Unfortunately, the authors point out, as does Schoenfeld, this characterization of instructional environments stands in sharp contrast to the ways most classrooms are organized and the methods by which most teachers assess student reasoning. The authors write of the typical math classroom, "*Doing* mathematics means following rules laid down by the teacher; *knowing* mathematics means remembering and applying the correct rule when the teacher asks a question, and mathematical *truth* is determined only when the answer is ratified by the teacher" (457). It is clear that

classrooms organized similarly to the typical classroom described by the authors are unable to provide the conditions necessary for the development of students' capacity to think and reason mathematically. The authors conclude that we must work with teachers to provide a learning environment characterized by reasoning, sense making, and discourse. This work helps us understand that enhanced instruction is a means of building students' capacity for mathematical thinking and reasoning. When students are provided with opportunities to engage in thinking, reasoning, and sense making in the mathematics classroom, it should lead to a deeper understanding of mathematics as well as the ability to demonstrate complex problem solving, reasoning, and communication skills on assessments of learning outcomes.

Hiebert and Wearne (1993) originally tested this very finding in second-grade math classes. Their research contributed to the Stein et al. study (1996) and supports the conclusions the authors drew from their research. Hiebert and Wearne set out to investigate the relationships between teaching and learning in twelve weeks of instruction in place value and multidigit addition and subtraction. They found that students in classrooms that emphasized constructing relationships rather than practicing proscribed procedures had greater test gains. Among the alternative approaches were assessments where the second-grade students received fewer problems, spent longer amounts of time with each problem, were asked verbal questions requesting them to describe and explain alternative strategies, and talked more using longer responses to explain their reasoning. They found that these students showed higher levels of performance at the end of the school year on most types of assessment items in addition to the alternative ones. Hiebert and Wearne, like Stein and her colleagues, conclude that their results suggest the

relationships between teaching and learning are a function of the instructional environment and that different relationships emerge in alternative-approach classrooms as compared with classrooms with traditional approaches. Hiebert and Wearne's work deepens the extant body with the conclusions that classroom tasks and classroom discourse define important links between teaching and learning. Their robust evidence supports their hypothesis that connections between teaching and learning are mediated by learning tasks and classroom discourse. However, the notion of mediation is not simple; the way in which tasks, which may induce different kinds of learning based on the cognitive processes that they require, and discourse connect teaching and learning in one classroom may be different than the way they connect in another, and this leads to important future research implications.

Implications

Research Implications

Within the assessment body, the implications tend toward developing future research agendas devoted to testing the validity of classroom-based formative assessments and specifically answering the empirical question best posed by Rea-Dickins and Gardner, "How is reliability achieved in observation driven assessments?" (Harlen, 2007a; Rea-Dickens, 2000). However, Black and Wiliam (1998) offer a more immediate implication that is related to the policy levers of accountability and the missing focus on helping teachers with the task of assessment. They conclude that, according to TIMSS (the Trends in International Mathematics and Science Study), "A focus on standards and accountability

that ignores the processes of teaching and learning in classrooms will not provide the direction that teachers need in their quest to improve" (Black, 1998). Their work shapes future research in the hopes of influencing future reform initiatives to change their aim to giving direct support to classroom teachers. They call into direct question the pervasive and harmful assumption held by policy makers that assessment is important solely for establishing a competitive market through education. They argue now for the need to move inside the "black box" of the classroom and explore the potential for assessment to raise standards as an integrated part of each pupil's learning. Thus, both policy environments and teacher education programs must begin with the locus of change within the classroom such that our overarching priority becomes effecting change within the daily classroom rather than continuing to measure the inputs and outputs and the black box. The input-output model may be helpful but it is not adequate to raise student achievement.

Across the empirical works reviewed, it's evident that there is a knowledge base of formative assessment that is theoretically framed and what is needed is a richer experiential frame. It seems that it is incumbent upon researchers in the United States to contribute to the field, as the studies reviewed have all been from researchers across the United Kingdom. Perhaps an additional empirical question must be asked about why the center of this research is currently located in Europe. What are the conditions that promote or constrain this line of research in the United States?

Further, while we have learned that the extant body on numerical proficiency and assessment concludes that we must work with teachers to provide a learning environment characterized by problem solving, sense making, and discourse, we must still undertake a

research agenda that includes closer examination of *how* such environments actually lead to the desired student outcomes. Empirical questions might include:

- Are authentic opportunities for students to think and reason mathematically created when teachers use assessments that demand explanation and justification?
- What kinds of thinking processes do these types of assessments set in motion?
- How does instruction that is designed to deepen mathematical reasoning relate to different learning outcomes?

Practitioner Implications

Rea-Dickins and Gardner (2000) conclude that a key implication of their research is for teacher development. They argue that whether it is within preservice training, formal training, or informal teacher-action projects, teachers need the opportunity to acquire skills in assessing, specifically in observation-based assessments because they capture a complexity of issues that are currently not taught or promoted.

Similarly, Harlen (2007a) finds that if teachers are not fully versed in administering and evaluating formative assessments, students may be unable to use feedback to improve their work or reveal their understandings, or misunderstandings, which are key to increasing student achievement. Harlen cites an example, which Wiliam's (n.d.) research also supports, from research that uncovered that teachers who gave marks with their comments on formative assessments had students who made less progress than those who received only targeted comments as feedback. Wiliam states, "In other words, if you write

careful diagnostic comments on a student's work, and then put a score grade on it, you are wasting your time" (1056). Harlan and Wiliam both argue that if a teacher does not know how to supply guidance for students to improve their work, the teacher will be unable to further learning because students will be unable to incorporate the feedback and improve their work. Wiliam concludes that this is a matter of mechanical or procedural understanding of assessment rather than purposeful understanding of formative assessment.

However, it is perhaps Wiliam's work that carries the greatest and most significant implications for practitioners in terms of the direct impact for improving student learning. Wiliam's findings on questioning and feedback across assessment types have the power to change the course of a student's mathematical trajectory. Wiliam (n.d.) finds that most teachers' questioning techniques establish only that the students' responses demonstrate that their mathematical ideas fit within the limitations of the questions asked. The questions, then, do not elicit student understanding. This finding has far-reaching impact for preservice training, professional development, and product development. To illustrate this impact, William developed the "window into thinking" rule for question and item development. The formative assessments we build and train teachers to implement will necessarily reveal students' misconceptions if they are to inform instruction and improve student learning.

Black and Wiliam (1998) charge us with the greatest task for practitioner research: developing answers and systems to support the research-action question, Is there evidence about *how* to improve instruction through formative assessment? The authors argue that for an assessment to function formatively, educators must use the results to actually adjust

instruction, so the most critical aspect of any reform initiative must be supporting the way teachers make such adjustments. Their work makes it clear that truly significant learning gains lie within our grasp as long as we are willing to make key classroom changes, especially, as Harlen (2007a) and Wiliam (2007) also discuss, those involving feedback between teacher and student. Black and Wiliam (1998) leave us with the caution, though, that teachers will not make these necessary changes in formative assessment practices if the ideas are only general principles and the task of translating them into everyday practice is left up to the teachers themselves.

The improvement of student learning through formative assessment is neither a quick fix nor a simple matter of identifying a typology of various assessments. If the substantial gains found in the research are to be secured, sustained programs of professional development and support must be implemented to show teachers when, why, and how to utilize the evidence gained from any assessment, formative or otherwise (Black, 1998). Black and Wiliam propose a four-point plan for this implementation:

1. Set up small numbers of model schools so teachers can learn from living examples, or appropriations.
2. Frequently and actively disseminate teachers' efforts and successes.
3. Reduce the obstacles between external summative tests and formative assessment practices.
4. Combine empirical and practitioner research that clearly outlines the actual classroom methods used in formative assessment.

Student achievement can be raised only by changes that are put into direct effect in the classroom. Black and Wiliam conclude that there is a robust body of evidence that the skillful use of formative assessment can raise achievement in a way that no other strategy can match.

Implications for Supporting Change in the Field: Human Capital and Product Development

Joyner and Muri (2011) conclude that the most likely ways to improve mathematics instruction are for teachers to shift their lessons and assessments to understand student reasoning and sense making and to use that knowledge to make instructional decisions. Joyner and Muri do not argue for a need to create a continuum of assessment types but rather argue for teachers to analyze students' reasoning in any assessment used and make alterations in their instruction based on the students' understandings or misunderstandings. The National Council of Teachers of Mathematics (1995) adds that assessment should support the learning of mathematics and provide useful, readily applicable information to both students and teachers.

Harlen's (2007a) research supports both of these conclusions, stating that the overall purpose of formative assessment is positive impact on teaching and learning and cautioning against the negative impacts that arise from formative assessments when teachers follow procedures mechanically without understanding their purpose. Harlen goes on to conclude that the running costs of formative assessment are zero. He states, "Once formative assessment methods are implemented, class time is used differently; attention is focused on developing students' understanding. . . . [T]here are upfront costs in

providing the professional development required, teachers need help in the form of descriptions of progression in various aspects of learning"(Harlen, 2007a).

Teachers require professional development not only to understand the purpose of formative assessment but to learn to collaboratively develop the key elements of formative assessment, such as Wiliam's "window into thinking" questions, in the style of the Japanese lesson study (Wiliam, 2004). Wiliam makes it clear to the education community that the success of any assessment lies in the question asked. Further, Wiliam concludes that formative assessments designed to increase student performance must include appropriate feedback opportunities and reflective assessment. Wiliam's conception of formative assessment shifts the focus from what teachers do to what students learn.

Schoenfeld's work (1992) supports Wiliam's call that instructional and assessment tools must support teachers in their problem-solving goals rather than, for example, fulfill the content-coverage routine of pencil-to-paper timed tests of rote mathematical memorization. Boesen, Lithner, and Palm's findings (2010) further reinforce these implications when the authors conclude that teacher-made assessments are often based on imitative reasoning that matches the instruction in the classroom, limiting the high-level problem solving we are intending for our students. They charge teachers and test developers alike to design assessment items that do not match the textbook and do not rely solely on student writing.

Schoenfeld goes on to argue that assessment may "be the single most potent systemic force in motivating change" (1992, 336). He supports his claim with arguments from the California Department of Education's *Everybody Counts*, which states, "What is tested gets taught. Tests must measure what is most important in mathematics" (1989,69).

Current state tests favor standardized, multiple-choice formats that deal with only a very small amount of sense making and mathematical reasoning. Schoenfeld calls for the development of appropriate assessment measures that answer the following questions:

- What kinds of information can be gleaned from open-ended questions?
- What kinds of scoring procedures are informative to both those who do the assessing and those who are being tested?
- What kinds of questions can assess students' fluency at generating a range of approaches to deal with difficult problems?

Discussion

Underlying the stated goals of this literature review is the goal of deepening the knowledge of policy makers and practitioners alike to enable them to enhance students' mathematical reasoning to help all students become numerically proficient. While the literature review opened by seeking clarity of terms, it became apparent through the analysis that the definition of the term matters less than the purpose for which the assessment, and the evidence from the assessment, is being used.

With increased emphasis being placed not only on students' capacity to understand the content of mathematics but also on their capacity to do mathematics, it is incumbent that research develop an agenda for policy makers to provide the professional development and assessment tools for teachers to heed this call. The researchers reviewed

here would likely all agree that classrooms should be communities where mathematical reasoning and sense making are practiced and environments where students are encouraged to discuss their problem-solving strategies with their teacher and each other, where intellectual risk taking is encouraged, and where student exploration of mathematical ideas is nurtured and respected. Stein, Grover, and Henningsen (1996) remind us that if students are not being set on the correct cognitive track during instruction, there is little reason to expect that scores on assessments will reflect enhanced understanding or the increased ability to reason and problem solve.

Across the research reviewed herein, it is clear that to understand what happens in classrooms, the educational community must understand how teaching and learning are connected. Future agendas will continue to explore the many variables inside the black box of instruction but one correlation that has been proved through the studies reviewed within is most succinctly summarized by Dylan Wiliam: "assessment is the bridge between teaching and learning" (2004, speech page 3). The authors of the CCSS suggest a concrete direction for our next steps when they remind policy makers, developers, and practitioners that asking a student to show his or her reasoning means asking a teacher to assess student reasoning (NGA Center and CCSSO 2010). With this imperative, then, teachers must be taught and must deeply grasp what mathematical reasoning looks and sounds like in the classroom. Only then can assessment fulfill its potential role as the most potent force in systemic education reform, as stated by Schoenfeld (1992).

References

Partnership for Assessment of Readiness for College and Careers (PARCC), P. f. A. o. R. f. C. a. C. (2012). Retrieved February 12, 2012, from www.parcconline.org

Black, P. J., and Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Boesen, J., Lithner, J., & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning studeents use. *Educational Studies in Mathematics, 75*(4), 89-105.

Burkhardt, H., Groves, S., Schoenfeld, A., & Stacey, K. (1988). *Problem Solving: A world view.* Paper presented at the Problem Solving Theme Group, ICME 5, London.

CCSSM. (2011). Retrieved March 21, 2012, from http://www.corestandards.org/the-standards/mathematics/introduction/standards-for-mathematical-practice/

Genesee, F., and Upshur, J.A. (1996). *Classroom-based evaluation in second language education*. New York: Cambridge University Press.

Harlen, W. (2007a). Criteria for evaluating systems for student assessment. *Studies in Educational Evaluation, 33*, 15-28.

Harlen, W. (2007b). *The quality of learning: Assessment alternatives for primary education*. Cambridge: University of Cambridge Faculty of Education.

Henningsen, M., & Stein, M.K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking. *Journal for Research in Mathematics Education, 28*(5), 524-549.

Joyner, J., and Muri, M. (2011). *INFORMative Assessment*. Sausalito: Math Solutions.

Lampert, M. (2003). *Teaching problems and the problems of teaching*. New Haven, CT: Yale University Press.

Levin, D., Fletcher, G., and Chau, Y. (2011). *Technology requirements for large-scale computer-based and online assessment: Current status and issues*.

Mathematics, N. C. o. T. o. (1995). *Assessment Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Merriam-Webster. (Ed.) (1997) (10th ed.). Springfield, MA: Merriam-Webster.

National Research Council (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington, DC: National Academy Press.

Rea-Dickens, P., and Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing, 17*(2), 215-243.

Schoenfeld, A. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for Research on Mathematics Teaching and Learning* (pp. 334-370). New York: MacMillan.

SMARTER Balanced Assessment Consortium. (2012). Retrieved February 15, 2012, from www.smartbalnaced.org

Stein, M. K., & Lane, S. (1996). Instructional tasks and the developmemt of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation, 2*(1), 50-80.

Stein, M. K., Grover, B.W., & Henningsen, M. (1996). Building student capacity for mathemtical thinking and reasoning: An analysis of mathemtical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455-488.

Wiliam, D. (2004). *Making Mathematics Vital*. Paper presented at the The Australian Association of Mathematics Teachers.

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1051-1089). Charlotte: Information Age Publishing.

http://www.parcconline.org/.%5BThishttp://www.smarterbalanced.org/.%5BThis