IDENTIFYING FORMATIVE ASSESSMENT IN CLASSROOM INSTRUCTION: CREATING AN INSTRUMENT TO OBSERVE USE OF FORMATIVE ASSESSMENT IN PRACTICE

by

Steven G. Oswalt

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Education in Curriculum and Instruction

Boise State University

December 2013

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the dissertation submitted by

Steven G. Oswalt

Dissertation Title: Identifying Formative Assessment in Classroom Instruction: Creating an Instrument to Observe Use of Formative Assessment in Practice

Date of Final Oral Examination: 20 August 2013

The following individuals read and discussed the dissertation submitted by student Steven G. Oswalt, and they evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Keith Thiede, Ph.D.	Co-Chair, Supervisory Committee
Michele Carney, Ph.D.	Co-Chair, Supervisory Committee
Richard Osguthorpe, Ph.D.	Member, Supervisory Committee
Jonathan Brendefur, Ph.D.	Member, Supervisory Committee

The final reading approval of the dissertation was granted by Keith Thiede, Ph.D., Co-Chair of the Supervisory Committee, and Michele Carney, Ph.D., Co-Chair of the Supervisory Committee. The dissertation was approved for the Graduate College by John R. Pelton, Ph.D., Dean of the Graduate College.

DEDICATION

This dissertation is dedicated to my wife Rhonda and my daughters Kimberly and Katey.

I love you very much.

ACKNOWLEDGEMENTS

I would like to express my appreciation and gratitude to all of those who have been a part of this process. While space does not permit me to mention all who have provided help and encouragement along the way, I would be remiss if I failed to mention some of those who have made this possible.

To my committee members: Dr. Keith Thiede, Dr. Michele Carney, Dr. Rich Osguthorpe, and Dr. Jonathan Brendefur. Thank you for all the time you have invested in me and in this dissertation. You have been patient, understanding, challenging, and insightful. I have never doubted that my success was your goal.

To my two fellow students and co-raters in this study: Amanda Bremner and Susan Woodard. Thank you. Without your hours of videotaping and observing classroom instruction, I could not have completed this study.

To all the BSU teachers and students who have been a part of my doctoral experience: You have all contributed to my journey in ways that you may never know, so thank you.

Finally, and most of all, to my wife Rhonda and my daughters Kimberly and Katey: Thank you for your patience, your encouragement, and your constant love.

Soli Deo Gloria

ABSTRACT

Is formative assessment observable in practice? Substantial claims have been made regarding the influence of formative assessment on student learning. However, if researchers cannot be confident whether and to what degree formative assessment is present in instruction, then how can they make claims with confidence regarding the efficacy of formative assessment? If it is uncertain whether and to what degree formative assessment is being used in practice, then any claims regarding its influence are difficult to support. This study aims to provide a vehicle through which researchers can make stronger, more substantiated reports about the presence and impact of formative assessment in classroom instruction. The ability to visually distinguish formative assessment during instruction would enable researchers to make such reports; therefore, this dissertation finds an appropriate method for identifying the presence of formative assessment to be an observational instrument.

In this study, a Formative Assessment Observational Instrument was developed for identifying formative assessment use in classroom instruction. The instrument was constructed around five components of formative assessment: understood learning targets, monitoring student learning, feedback, self-assessment, and peer assessment. Each component contained 3-5 scales for observation, each rated on a 1-5 Likert-type scale, totaling 20 items. Pairs of trained raters used the instrument to observe and rate 47 elementary mathematics instructional sessions, evenly divided between 16 teachers, of up to 30 minutes in length. Using the results of these observations, the instrument was

evaluated on the basis of reliability across time, reliability across raters, and reliability of scale. Based on these criteria, the instrument was found to be reliable for the purpose of identifying formative assessment in practice, and the instrument identified varying degrees of formative assessment use in terms of item, scale, and teacher.

As a result of examining the literature on formative assessment and utilizing this instrument in practice, it was proposed that in order for formative assessment to become a more quantifiable factor in researching influences on student learning, a narrowing and focusing of its definition was in order. Consequently, a more focused definition of formative assessment was suggested, defining formative assessment as a dynamic interchange between teacher and student in which instruction is adapted continuously based on student learning status. This definition narrowed formative assessment to what happens within instruction, calling for outside of classroom uses of assessment to be treated as separate factors in instruction. The definition also affirmed the first three components of formative assessment as comprising the essential nature of formative assessment. It distinguished self-assessment and peer assessment as methods for accomplishing those components, rather than as components themselves.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Background

Formative assessment has become a matter of much discussion in education, particularly since Black and Wiliam (1998) published their findings in their widely-cited article "Inside the Black Box: Raising Standards Through Classroom Assessment." Researchers claim that the use of formative assessment has a positive effect on student achievement (Black & Wiliam, 1998; Popham, 2008; Wiliam, Lee, Harrison, & Black, 2004). The research base for this claim, however, is not extensive, and a substantial amount of it rests on questionable research methodology (Bennett, 2011; Dunn & Mulvenon, 2009). The difficulty in showing the link between formative assessment use and student achievement (thereby weakening research claims) may result from the difficulty of distinguishing the elements of formative assessment from other aspects of teaching. That is, making a clear distinction between the use of formative assessment and other foundational teaching practices can be challenging, which may explain why much of the writing done to date on formative assessment has been based in theoretical discussions rather than empirical research.

It is crucial, however, that decisions regarding proposed educational programs and practices be based on empirical research and not simply theoretical discussions. As Robert Slavin (2008) wrote:

> Throughout the history of education, the adoption of instructional programs and practices has been driven more by ideology, faddism, politics, and marketing than

> by evidence. For example, educators choose textbooks, computer software, and professional development programs with little regard for the extent of their research support. Evidence of effectiveness of educational programs is often cited to justify decisions already made or opinions already held, but educational program adoption more often follows the pendulum swing of fashion, in which practices become widespread despite limited evidentiary support and then fade away regardless of the findings of evaluations. (p. 5)

If formative assessment truly supports student learning, then it is critical that we show that it does empirically so that it does not fade away. In order to do this, we must develop a method to measure its use in classroom instruction. Additionally, if we are to understand what elements of formative assessment are most influential, most neglected, and most misunderstood, we must be able to observe those elements in actual classroom instruction.

Statement of the Problem

Is formative assessment observable in practice? If we are to ascertain whether and to what degree formative assessment is occurring in classroom instruction, we must have a method to observe its use. If we are to ascertain whether formative assessment truly makes a significant difference in student learning, we must have a method to observe its use. Developing such a method could be important for enhancing formative assessment use because it has been noted that reliable observational tools are essential for providing teachers with meaningful feedback on their classroom practice and for understanding patterns of implementation that can direct professional development (Baker, Gersten, Haager, & Dingle, 2006).

While formative assessment practices can be found embedded in both teacher training and evaluation, there remains an absence of an instrument/method specifically intended to identify its use. In light of the potential impact of formative assessment on student learning, it is important to develop such a method for it to be observed. The purpose of this endeavor is to determine whether formative assessment can be observed in practice through the creation and implementation of an observational instrument designed for that purpose.

Of course, whether it is observable is an unnecessary question unless the use of formative assessment actually makes a difference for students. Advocates of formative assessment claim that it does make a difference in classroom instruction, but the question remains of how much the use of formative assessment truly affects what matters most – student learning. This is the key question regarding formative assessment, for as Harlen (2007) said, "Formative assessment has a single clear purpose: that of helping learning and teaching. If it does not serve this purpose it is not, by definition, formative" (p. 19). While the research on the effectiveness of formative assessment has been hopeful, the difficulty in distinguishing it as a separate variable for study has made the results difficult to interpret. Black and Wiliam (1998) proposed that formative assessment can be a powerful strategy for increasing achievement for all students, even those students who might be low achievers. Others have echoed the idea that students are best served through an educational approach that uses assessment to improve rather than simply report student achievement (Marzano, 2006; Stiggins, 2005; Wiliam et al., 2004). In *Raising Student Achievement through Rapid Assessment and Test Reform*, Stuart Yeh (2006) presented evidence from a study suggesting that rapid assessment (systems that

test students 2 to 5 times per week in math and reading and provide rapid feedback of the results to students and teachers) is at least eight times as effective as a 10% increase in per pupil expenditure, seven times as effective as charter schools or vouchers, and 14 times as effective as accountability alone. In another study, Yeh (2008) claimed that results indicate rapid assessment represents a much more cost-effective approach than Comprehensive School Reform programs, class size reduction, or high quality preschool.

Some critical voices, however, have called into question the research claims of formative assessment proponents (Bennett, 2011). The claims of Black and Wiliam (1998) have been critiqued specifically in reference to methodological issues with the studies they examined, with the resulting conclusion that those studies do not support the reported effect sizes of around .70 for formative assessment impact (Dunn & Mulvenon, 2009; Kingston & Nash, 2011). In evaluating the effect of formative assessment, Kingston and Nash (2011) found that despite many hundreds of articles written on formative assessment, they were able to find only 42 usable effect sizes from 1988 to the present. By using a random effects meta-analytic approach to analyze them, they found that the weighted mean effect size was .20 and the median of the observed effect sizes was .25. Despite their critique, however, they stated that "results, though, do indicate formative assessment can be a significant and readily achievable source of improved student learning" (Kingston & Nash, 2011, p. 33). Likewise, despite Dunn and Mulvenon's (2009) extensive critique of the research methodologies used in examining the effect of formative assessment, they made clear that their purpose was not to deny the positive effect of formative assessment but rather to instigate continued research into it. They wrote:

> The research discussed in the Black and Wiliam's (1998) review and the other research discussed here does provide some support for the impact of formative assessment on student achievement. However, it provides greater support for the need to conduct research in which more efficient methodologies and designs will lead to more conclusive results and understanding of the impact of formative assessment and evaluation on student achievement. (p. 9)

One of the strongest claims for connecting formative assessment to student learning has been made by John Hattie. In a synthesis of over 800 meta-analyses relating to student achievement, John Hattie (2009) found that out of 138 influences on student achievement (including such influences as teacher-student relationships, home environment, socio-economic status, and class size), the third most positive influence on student achievement was formative evaluation, with an effect size of 0.9. His finding makes such a strong statement of support for the efficacy of formative assessment that we must examine his work, specifically the nature and limitations of his methodology and the synonymity of his term *formative evaluation* with the term *formative assessment*.

While some may differ on a strict definition of a meta-analysis, Gliner, Morgan, and Harmon (2003) define it in a manner fitting for Hattie's approach: "a research synthesis that uses a quantitative measure, effect size, to indicate the strength of relationship between the treatments and dependent measures of studies making up that synthesis" (p. 1376). Such meta-analyses have received criticism, primarily for the potential error and bias that may result from combining studies. These criticisms include the *apples and oranges* problem (differences between studies), the *garbage in-garbage out* problem (differences in methodological quality between studies), the *a priori* problem

(inclusion/exclusion of specific studies), and the *file-drawer* problem (the issue of publication bias towards studies with significant results) (Shelby & Vaske, 2008). Hattie went a step beyond meta-analysis by creating a synthesis, or meta-meta-analysis, of more than 800 meta-analyses (encompassing 52,637 studies and providing 146,142 effect sizes) about influences on learning. While he acknowledged the potential problems inherent in meta-analyses, Hattie (2009) responded that "the generalizability of the overall effect is an empirical issue, and…there are far fewer moderators than are commonly thought" (p. 10). He reported the impact, therefore, of various influences on learning by using a fixed effect size model. He also argues for isolating specific variables across diverse studies as a counter to the *apples and oranges* objection and for simply recognizing design quality as one moderator of conclusions as a counter to the *garbage in-garbage out* objection.

Did Hattie (2009) use the terms *formative evaluation* and *formative assessment* synonymously? For the most part, the answer is yes. To best understand his use of formative evaluation, however, we must first understand his usage of the term *feedback* because, for him, feedback contains a strong flavor of formative assessment. Hattie moves away from the idea of feedback as something exclusively provided by teachers to students. Instead, he wrote:

> It was only when I discovered that feedback was most powerful when it is from the *student to the teacher* that I started to understand it better. When teachers seek, or at least are open to, feedback from students as to what students know, what they understand, where they make errors, when they have misconceptions,

> when they are not engaged – then teaching and learning can be synchronized and powerful. Feedback to teachers helps make learning visible. (p. 173)

As Hattie continued to discuss feedback, he made clear that he intended it to be used diagnostically by teachers in order to guide their instruction. He explicitly stated:

> [A] major argument throughout this book is the power of feedback to teachers on what is happening in their classroom so that they can ascertain "How am I going?" in achieving the learning intentions they have set for their students, such that they can then decide "Where to next?" for the students. Formative evaluation provides one such form of feedback. (p. 181)

When used in this regard, feedback becomes a diagnostic method of continually monitoring student learning in light of established learning intentions in order to adapt instruction accordingly – which is the essence of formative assessment.

The importance of teachers receiving feedback from students is acknowledged as a major theme in Hattie's (2009) book, and he uses the term *formative evaluation* to describe this particular process of teachers continually evaluating the effects of their teaching, specifically in regard to student learning progress. This was so important to him that Hattie stated the "major message is for teachers to pay attention to the formative effects of their teaching, as it is these attributes of seeking formative evaluation of the effects (intended and unintended) of their programs that makes for excellence in teaching" (p. 181). Hattie's use of formative evaluation and formative assessment can be fairly equated and, in addition, his use of feedback can also be closely related to the process of formative assessment.

It may well be, as claimed by researchers, that one of the most powerful factors in student learning can be found in formative assessment. In the realities of student learning, formative assessment may be a pathway to success both in summative assessment and in post-assessment retention of information. Formative assessment may hold power in positively influencing both the teacher's approach to instruction and the student's approach to learning, an approach wherein the students learn not only the material at hand but also learn the approaches to learning that work best for them. Formative assessment may hold such power and potential; however, without the ability to observe it in practice, how will we know?

Research Question

The research question for this study is: Is formative assessment observable in practice?

Research Hypothesis

The research hypothesis for this study is that formative assessment is observable in practice.

CHAPTER TWO: REVIEW OF LITERATURE

Introduction

This review of literature consists of three parts. The first part will provide an overview of formative assessment. This will include an examination of the nature of formative assessment, definitional difficulties in conceptualizing formative assessment, and the relationship of formative assessment with content knowledge and depth of knowledge. The second part will examine an operational definition of formative assessment. Based on that chosen operational framework, it will then examine five specific formative assessment strategies and their use in classroom instruction. The third part of this review of literature will examine existing observational instruments that include evaluations of formative assessment in classroom instruction. The nature and purpose of these instruments will be examined, as well as the format for their implementation.

Part 1: An Overview of Formative Assessment

The Nature of Formative Assessment

The word "assessment" conveys a sense of high stakes. Outside of the educational world, the word "assessment" is most often connected with financial matters. Its definition has included "the act of assessing, appraisal; evaluation," "an official valuation of property for the purpose of levying a tax," and "an amount assessed as payable" ("assessment," n.d.). When assessment is applied in the world of education, the

stakes are no less high. This became especially true with the advent of the No Child Left Behind (NCLB) legislation of 2001 and its high-stakes accountability systems based on standardized testing (Sunderman & Kim, 2007). Due to the failures of NCLB (Rothstein, Jacobsen, & Wilder, 2008), assessment has come to be viewed negatively by many. This is unfortunate because assessment, which in education may be defined as "a measurement of the learner's achievement and progress in a learning process" (Gikandi, Morrow, & Davis, 2011, p. 57), is a crucial aspect of the educational process.

Assessment can be used for different purposes in education. This is reflected in Wiggins and McTighe's (2005) definition of assessment as "the act of determining the extent to which the desired results are on the way to being achieved and to what extent they have been achieved" (p. 6). Even in this simple definition, two different purposes are identified: looking at results as completed and looking at results as in process. Assessment is commonly divided into two types that relate to these two purposes: summative and formative assessment (Chappuis & Chappuis, 2008). Summative assessment is designed primarily to document what students know, that is, what has been achieved in the instructional endeavor. This is how assessment has been most often used, as the "processes of evaluating the effectiveness of a sequence of instructional activities when the sequence was completed" (Wiliam, 2011, p. 3). Formative assessment is designed primarily to deliver information during the instructional process to help make decisions about what actions will promote further learning (Chappuis & Chappuis, 2008). This distinction has been referred to by Stiggins (2002) as assessment of learning (i.e., summative) versus assessment for learning (i.e., formative).

It should be noted that Newton (2007) suggested three purposes for the use of assessment: judgment, decision, and impact. The *judgment* level relates most closely to summative assessment as it seeks to determine the extent of completed learning. The *decision* level relates most closely to formative assessment as it seeks to use information from assessment to make decisions regarding future instruction. The *impact* level relates to the purpose that assessment plays in the affective component of a student's learning. Newton's purposes of assessment mesh well with the model proposed by Bennett (2010), which characterizes assessment as *of*, *for*, and *as* learning. While these three uses of assessment are of interest and potential significance, the focus of this study will specifically be on the formative use of assessment (i.e., assessment for learning).

If we are to focus on the formative use of assessment, it is crucial the understanding of its nature and purpose be clear. So, what is formative assessment? A consensus definition of formative assessment has proven elusive (Dunn & Mulvenon, 2009). Educational researchers have written about the concept of formative assessment for decades, although they have used different terms to discuss it. In 1967, the term *formative evaluation* was used to describe "feedback on the basis of which he [an instructor] again produces revisions" (Tyler, Gagné, & Scriven, 1967, p. 43). The term *formative assessment* may have been coined by Bloom (1969) in applying the distinction made by Scriven (1967) between formative and summative program evaluation to the evaluation of individual students (Thompson & Wiliam 2008). Other terms have included *formative observation* (Bloom, Hastings, & Madaus, 1971), *transformative assessment* (Popham, 2008), and *rapid assessment* (Yeh, 2006). Tomlinson (2008) used the term *informative assessment* to describe this approach to support student learning,

albeit in an informal sense, observing that "giving students feedback seemed to be more productive than giving them grades" (p. 10).

Definitional Difficulties

One of the issues in delineating the nature and purpose of formative assessment has been determining how narrowly or broadly to define it. Black and Wiliam (1998), whose work increased attention on formative assessment, defined formative assessment broadly as "all those activities undertaken by teachers – and by their students in assessing themselves – that provide information to be used as feedback to modify teaching and learning activities" (p. 140). This broad, all-encompassing definition was echoed by Stiggins (2005) in his description of formative assessment as *assessment for learning*. W. James Popham (2008) focused more on the role of intentional planning, defining formative assessment as "a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning tactics" (p. 6).

Another way of considering how narrowly or broadly to define formative assessment is by asking whether it should be considered an instrument or a process. Is it a product, a process, or a package to be bought from curriculum and assessment vendors (Pinchok, Brandt, & Learning Point, 2009)? Bennett (2011) addressed this question of whether formative assessment should be considered an instrument or a process, described the alternative perspectives, and ultimately argued for an integration of the two. He stated that "formative assessment then might be best conceived as neither a test nor a process, but some thoughtful integration of process *and* purposefully designed methodology or instrumentation" (Bennett, 2011, p. 7). However, it would seem that the

attempt to integrate process and instrumentation ignores the fundamental nature of formative assessment, which is the dynamic interchange between teacher and student in which there is an ongoing iterative process of evaluation and adaptation by the teacher and student. While instrumentation may certainly be used as part of the process, it is not the process itself. Though the usage of the term remains somewhat amorphous and some assessment vendors may still disagree, formative assessment is best understood not as a particular assessment tool but rather as a matter of the uses to which assessment data are put (Andrade, 2010).

Another question to consider in delineating the specific nature of formative assessment relates to time. When discussing formative assessment, how long is the period of time between diagnosing the status of student learning and adapting instruction? To help clarify, Wiliam and Thompson (2008) created a typology of formative assessment distinguishing between long-cycle (across marking periods, quarters, semesters, or years), medium-cycle (within and between instructional units), and short-cycle (within and between lessons) types of formative assessment. The short-cycle type of formative assessment is reflected in the professional development program *Keeping Learning on Track*® (KLT) by ETS, which frames its "big idea" definition of formative assessment as "students and teachers using evidence of learning to adapt teaching and learning to meet immediate learning needs minute-to-minute and day-by-day" (Thompson & Wiliam 2008). While each of these time frames bears separate investigation, seeking to identify formative assessment use in instruction calls for focusing on what is observable, which is the short-cycle nature of formative assessment (i.e., the moment-by-moment use of formative assessment within a single instructional

session). Perhaps more importantly, it is that dynamic usage that fits most closely with the concept of formative assessment as a process used by teachers and students during instruction.

A definition of formative assessment from 2007 highlights these essential characteristics of time and process. During 2006, the Council of Chief State School Officers (CCSSO) created a sub-entity, the Formative Assessment for Students and Teachers – State Collaborative on Assessment and Student Standards (FAST SCASS) to address formative assessment (Popham, 2008). Later that year, FAST SCASS held a meeting to determine a definition of formative assessment, publishing the following definition in 2007: "Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (CCSSO, 2012). Several elements of this definition are worthy of note and foreshadow key formative assessment strategies. This definition highlights that formative assessment is done in light of intended instructional outcomes, which relates to maintaining clear learning objectives. It highlights the role of feedback, which is central to all formative assessment. It highlights that formative assessment is a process involving both teachers *and* students, which relates to self-assessment and peer-assessment formative assessment strategies. And this definition makes clear that formative assessment is a process that happens during instruction…to improve students' achievement, thus focusing on a time frame of immediacy within a classroom.

A helpful term in considering the way in which formative assessment naturally integrates with classroom instruction is the term *informal formative assessment.* Jordan

and Putz (2004) distinguish formal, or documentary, assessment that includes the use of objective standards of measurement with informal assessments that center around the ongoing interpersonal evaluations that form a natural aspect of dynamic human exchanges, be they verbal or non-verbal. While they described the use of informal assessment in all social contexts, Ruiz-Primo (2011) applies this concept to the world of education, applying the construct of formative assessment to the everyday practices of teachers. She holds that assessment conversations happen every day in the classroom and that these conversations make students' thinking explicit, thus allowing their thinking to be examined, questioned, and shaped constructively. As opposed to formal, scheduled assessment tasks, she proposes the use of informal formative assessment as "the small-scale, frequent opportunities teachers have for collecting information about their students' progress towards the learning goals they have in mind" (Ruiz-Primo, 2011, p. 16). Thus, the idea of *informal formative assessment* carries with it the idea of embedding the practice of formative assessment naturally within the normal course of daily classroom life.

Content Knowledge and Depth of Knowledge in Formative Assessment

In defining and seeking to operationalize formative assessment, the question of content knowledge arises. Bennett's (2011) analysis of formative assessment raised the issue of whether formative assessment training and use is domain dependent or independent of domain. He made the point that the attempted use of formative assessment strategies by a teacher having inadequate domain knowledge will fail to support learning due to the teacher's inability to accurately assess learning in that domain and to adapt instruction in that domain to support the student's progress. He argued for

training in both domain knowledge and formative assessment, stating that "to be maximally effective, formative assessment requires the interaction of general principles, strategies, and techniques *with* reasonably deep cognitive-domain understanding" (Bennett, 2011, p. 15). While it is unlikely that anyone would argue against the importance of content knowledge in teaching, content knowledge alone does not an effective teacher make. It is the use of that content knowledge in supporting the learning of students that is the goal of instruction.

Perhaps the better question is whether the nature of formative assessment supports student learning on a deeper level of knowledge, at a conceptual rather than a more shallow procedural level, irrespective of the subject matter. Whether we are considering instruction in reading or in mathematics, this deeper knowledge is to be the goal. Hiebert and Lefevre (1986) discussed the distinction between procedural and conceptual knowledge in terms of mathematics education, but their understanding of these different types of knowledge also applies to other curricular subjects. Procedural knowledge deals with forms/symbols and rules, in mathematics those being numbers as symbols/forms and algorithms as rules. Theoretically, in reading, forms/symbols might equate to representing text at the word or propositional level (i.e., what Kintsch, 1988, referred to as the textbase). Conceptual knowledge deals with relationships between bits of knowledge and can be thought of as "a connected web of knowledge, a network in which the linking relationships are as prominent as the discrete pieces of information" (Hiebert & Lefevre, 1986, pp. 2-3). In reading, this conceptual knowledge might apply to comprehension, particularly the activation of background knowledge (Marzano, 2003; Moreillon, 2007)—Kintsch (1988) referred to this as the situation model of a text. Both

conceptual and procedural knowledge are needed, and Hiebert and Levefre (1986) found that the two types of knowledge may work in partnership to support the ultimate goal of helping students build deep, transferrable knowledge.

Regarding formative assessment, while it can certainly benefit students through observing and correcting procedural problems in learning, its greatest potential resides in its ability to bring students to deeper levels of understanding. Hattie (2009), who found formative assessment had the third most powerful influence on student learning, stated that "the major influences on achievement cross curriculum boundaries – the more important attribute is the balance of surface or deep understanding within each curriculum subject, which leads to conceptual clarity" (p. 35).

Conclusion

From this overview of formative assessment, it should be clear that there is a need for more clarity regarding its nature. While most researchers are in substantial agreement as to the fundamental nature of formative assessment, their vernacular is inconsistent and their individual emphases vary. This can lead to a lack of clarity regarding whether formative assessment is primarily a tool or a technique, whether it should be planned or spontaneous, whether it can be immediate or delayed or both, whether it should be driven primarily by students or by teachers, and how it can be distinguished from general pedagogical practices. No one doubts the existence of formative assessment, but in order to truly appreciate the power it may wield for learning, a more focused understanding of what it looks like in practice would be beneficial. In order to observe it, however, we must operationalize it into observable scales.

Part 2: An Operational Definition of Formative Assessment

If we are to examine and investigate the nature and impact of formative assessment, it is necessary to operationalize the construct of formative assessment into observable scales (Bennett, 2011). Operationally, therefore, what are the key scales of formative assessment? The FAST SCASS identified the following five attributes as critical features of formative assessment (McManus, 2008):

- Learning Progressions: Learning progressions should clearly articulate the sub-goals of the ultimate learning goal.
- Learning Goals and Criteria for Success: Learning goals and criteria for success should be clearly identified and communicated to students.
- Descriptive Feedback: Students should be provided with evidence-based feedback that is linked to the intended instructional outcomes and criteria for success.
- Self- and Peer-Assessment: Both self- and peer-assessment are important for providing students an opportunity to think metacognitively about their learning.
- Collaboration: A classroom culture in which teachers and students are partners in learning should be established.

In describing these five attributes, they cautioned that no one of them should be regarded as "a *sine qua non,* that is, an attribute without which the assessment would not be formative" (McManus, 2008, p. 4). They also made clear that the implementation of these attributes would depend on the particular instructional context, the teacher, and the students. While these attributes provide helpful guidance, they fail to highlight the central function of a teacher's monitoring of student learning progress.

A similar, but different, operational structure is used by ETS's professional development program *Keeping Learning on Track*® (KLT). While they also acknowledge that there are no one-size-fits-all tools or techniques, they employ five key strategies for correctly and effectively utilizing formative assessment, with the specific implementation of these strategies being determined by the classroom teacher (Bennett, 2011; Leahy, Lyon, Thompson, & Wiliam, 2005; Thompson & Wiliam 2008). These five key strategies are:

1. Clarifying and sharing learning intentions and criteria for success.
2. Engineering effective classroom discussions, questions, and learning tasks.
3. Providing feedback that moves learners forward.
4. Activating students as the owners of their own learning.
5. Activating students as instructional resources for one another.

These five strategies overlap substantially with the attributes of formative assessment from FAST SCASS, with the primary difference being the inclusion of activities designed to monitor student learning and the exclusion of the attributes of learning progressions and collaboration, elements logically encompassed by the other strategies. The five key strategies of KLT provide an operational structure that can be observable in classroom practices. Consequently, these five strategies should provide the best operational focus for the creation of an observational instrument to evaluate a teacher's use of formative assessment in a classroom. The fact that they form the basis for ETS's comprehensive professional development program *Keeping Learning on Track*® (KLT) only serves to support their usefulness (Wylie, 2008).

In order to understand the nature of each of these five strategies, I will discuss them individually and provide examples of how they might be utilized. Fortunately, instructional practices utilizing formative assessment have become common in education texts. For example, an internet search on "formative assessment" in the books section of Amazon.com returned 1,232 results (Amazon, 2013). Some texts explicitly provide examples for implementing these five specific operational strategies (Earl, 2013), some provide instruction on implementing formative assessment more generally (Heritage, 2010), and some embed formative assessment practices within the broader spectrum of assessment and instruction (Baldwin, Keating, & Bachman, 2006; Joyce, Weil, & Calhoun, 2011; Russell, Airasian, & Airasian, 2012; Wiggins & McTighe, 2005). Ideas and examples from such texts are included within the description of each of the five formative assessment strategies that follow.

Clarifying and Sharing Learning Targets and Criteria for Success

The first operational scale of formative assessment is the clear understanding of learning targets. This strategy of *clarifying and sharing learning intentions and criteria for success* means that students and teachers both clearly understand how success is defined. This scale of clear, shared understanding of learning targets serves as a foundation for all other scales of formative assessment because if teachers and students do not have this, then there is no basis for evaluating student progress towards them. The difference between desired results and current status has been described as the learning gap between what students know and what they need to know (Sadler, 1989). Sadler explained that a learner must:

(a) possess a concept of the *standard* (or goal, or reference level) being aimed for,

(b) compare the *actual* (or current) *level of performance* with the standard, and

(c) engage in appropriate *action* which leads to some closure of the gap. (p. 121)

The clear understanding of learning targets is crucial because it has been suggested that low achievement often comes simply because students do not know what is required of them (Black & Wiliam, 1998). The setting of clear learning targets (be they behavioral objectives or cognitive goals) is a well-established instructional foundation (Wiggins & McTighe, 2005). It is important to note that clarifying and sharing learning intentions does not simply involve posting objectives on a board, but it includes various ways that teachers can make transparent to students the criteria for their success (Leahy et al., 2005). It means coming to deeply understand characteristics of quality work, taking the time to help students see what quality work and performance look like so that the learning targets/standards are not a mystery to them (Pinchok et al., 2009).

Effective communication of learning targets between teachers to students will not happen without intentionality, and unfortunately, may not happen at all. Urdan (2004) conducted studies regarding student perceptions of classroom goal structures. In one study, he used observational and interview methodologies with 24 elementary and middle school students and their teachers from four classrooms. He found that goals were rarely explicitly discussed by teachers or students in the classroom and that the teachers often provided mixed and contradictory goal messages. He also found that students differed in their perception of and reaction to goal messages, partly according to age and achievement levels.

Perhaps the presence of such a disconnection between teacher and student understanding of learning goals can be understood through the work of Entwistle and Smith. In an examination into student learning outcomes as related to various theories of learning, Entwistle and Smith (2002) described an important distinction between target understanding and personal understanding. Target understanding is defined largely by the syllabus and the teacher.

> Target understanding is shown as originating in decisions taken by the curriculum designers about course specifications or examination syllabuses; these produce the formal target. Interpreting that target, teachers are influenced by their own knowledge and attitudes about the subject, and by their beliefs about teaching and learning. (Entwistle & Smith, 2002, p. 335)

Personal understanding, on the other hand, refers to the student's conceptualization of the task at hand, as influenced by factors such as how they perceive it and the background experiences the student brings to the task.

> The teacher's target is interpreted by the students through the filter of their existing knowledge and personal histories, including their attitudes, beliefs, and self-concepts. All of these affect their motivation and approach to studying within the classroom, their comprehension of the target, and their perception of the learning context. These three scales then influence the learning strategies, effort, and engagement that students show in carrying out the task, resulting in a personal understanding of the topic which is then evaluated by the teacher or examiner. (Entwistle & Smith, 2002, pp. 335-336)

An understanding of this distinction between target understanding and personal understanding is at the heart of the first formative assessment strategy. Without clear communication of expectations, whatever form that may take, and clear understanding of those expectations by students, teachers and students may unknowingly pursue divergent goals, compromise student success, and enhance the likelihood of frustration by both teachers and students.

In addition to increasing intrinsic motivation (Ames, 1992; Ames & Archer, 1988; Murayama & Elliot, 2009), research has shown that the discussion of criteria and exemplars in class, at least in university settings, can result in increased student understanding of standards and higher achievement (Hendry, Armstrong, & Bromberger, 2011). Rust, Price, and O'Donovan (2003) found with college students that a structured process involving a workshop and peer collaboration helped develop student understanding of assessment criteria and the assessment process, with a resultant significant increase ($p < 0.01$) in their achievement of a .6 effect size in one cohort and a .69 effect size in another cohort.

Various resources provide examples of practices that teachers may use for enacting this strategy. Such practices may include providing students with exemplars, examples of quality work that may be in the form of student work from another class or teacher-made mock-ups (Thompson & Wiliam 2008). A similar technique involves providing students with non-examples, which some have suggested is an even more powerful means of helping students understand learning intentions and criteria for success (Archer & Hughes, 2010; Marzano, 2003). Another possible technique for this strategy is to enlist students to create practice tests or test items as a method to gauge

their understanding of learning concepts and essential concepts (Chappuis & Stiggins, 2002).

It is important to remember that the strategy of *clarifying and sharing learning intentions and criteria for success* applies to the entire instructional session, not merely to its beginning. Margaret Heritage (2010) published a book that provided an extensive examination of formative assessment, attempting to bridge theory, research, and practice. She discussed actual classroom practices in formative assessment by teachers in Iowa, in Syracuse, New York, and in Los Angeles, California. In that book, she devoted an entire chapter to "The Drivers of Formative Assessment: Learning Goals and Success Criteria," in which she provided examples of how teachers may create learning goals comprehensible to students. In discussing the importance of teachers communicating the learning goals and criteria for success to students from the beginning of class, she reminded us that an important method of accomplishing that is for teachers to draw attention back to those criteria while teaching. One teacher she observed commented that "I can talk and write about plotting points on a coordinate grid using correct vocabulary" (Heritage, 2010, p. 54) as a way of helping students remember and understand the meaning of their success criteria, which then became the springboard for their group tasks. The clear communication and reminders regarding learning targets and criteria for success can support effective use of formative assessment.

Engineering Effective Classroom Discussions, Questions, and Learning Tasks

The second operational scale of formative assessment is the monitoring of student learning. This strategy of *engineering effective classroom discussions, questions, and learning tasks* focuses on the teacher's ability to diagnose the state of student learning on

an ongoing basis. This scale is interwoven throughout the formative assessment process, but must be separated for study. Bennett (2011) made the point that there are two key mechanisms involved in formative assessment: making inferences about student learning and using those inferences to adapt instruction. He observed "that distinction, between making evidence-based inferences and subsequently adapting instruction is crucial…because a failure in either step can reduce the effectiveness of formative assessment" (Bennett, 2011, p. 14). Thus, it is important that this strategy be separated for examination as one of the operational scales.

The breadth of techniques that can be used to monitor student learning through engineering effective classroom discussions, questions, and learning tasks is admittedly immense. One technique, however, is worthy of particular attention because of the manner in which it often permeates classroom instruction - questioning. Unfortunately, questioning in the classroom is too often done shallowly, narrowly, or ineffectively (Leahy et al., 2005; Pinchok et al., 2009). When used formatively, however, questioning can be for such purposes as eliciting information, probing thoughts and ideas, tapping into different types of knowledge, and instigating deeper levels of understanding (Ruiz-Primo, 2011). To help teachers improve their questioning skills, Walsh and Sattes (2005) proposed the simple acronym QUILT (Questioning and Understanding to Improve Learning and Thinking) as a starting point (as cited in Fisher & Frey, 2007, p. 37-41). Step 1 is to prepare the question, particularly in terms of the purpose and type of question to ask. Step 2 is to present the question, making clear how the question is to be answered and who is to answer it. Step 3 is to prompt student responses, providing adequate "wait time" (3-5 seconds) after asking the question, scaffolding the question for students who

struggle, and pausing after the answer so that students can think about it. Step 4 is to process student responses, providing appropriate feedback while continuing to utilize follow-up probes for elaboration on incorrect answers and expansion on correct answers. Step 5 is to reflect on the questioning process, identifying ways to improve future questioning and encourage the participation of students. While other areas of effective questioning skills could be considered, such as providing non-verbal support and developing authentic questions (Fisher & Frey, 2007), implementing the skill sets of QUILT would be of great use to teachers in diagnosing the state of their students' learning. As one teacher stated, "Good questioning is really about the ability to recognize when the quiet kid doesn't get it" (Volante & Beckett, 2011, p. 244).

The use of this formative assessment strategy relates to what Ruiz-Primo and Furtak (2007) described as assessment conversations that "permit teachers to recognize students' conceptions, mental models, strategies, language use, or communication skills, and allow them to use this information to guide instruction" (p. 60). Ruiz-Primo and Furtak (2007) conducted a study exploring teachers' informal formative assessment practices in three middle school science classrooms, utilizing a model that examined formative assessment as occurring in iterative ESRU cycles in which the teacher Elicits a question; the Student responds; the teacher Recognizes the student's response; and then Uses the information collected to support student learning. The first three parts of this ESRU cycle (eliciting information, student response, and recognition of response) track closely with the formative assessment strategy of monitoring student learning through engineering effective classroom discussions, questions, and learning tasks. Extensive video-taping of the teachers (30 lessons across the three teachers) involved in the study

revealed differences in the amount of assessment conversations each teacher used and in how much those conversations were aligned with the desired conception of informal assessment practices in the context of scientific inquiry. The teacher who most used assessment conversations aligned with the ESRU cycle scored significantly higher on three embedded assessments of student science learning: Graphing [$F_{(2, 69)}$=5.564, p=0.006, R^2=0.139], Predict-Observe-Explain [$F_{(2, 70)=}$28.939, p=0.000, R^2=0.453], and Prediction Question [$F_{(2, 51)}$=5.257, p=0.008, R^2=0.171]. There had been no significant difference between students on the pretest given before the experiment. While the generalizability of the study is limited by only involving three teachers, the results do support the idea that such formative practices may lead to improved student performance.

A teacher may find numerous examples of formative assessment techniques that support the strategy of engineering effective classroom discussions, questions, and tasks to monitor student learning (Fisher & Frey, 2007; Heritage, 2010; Thompson & Wiliam 2008; Wiggins & McTighe, 2005). In addition to verbal questioning techniques, a teacher may utilize written communications from students, such as an "entrance ticket" providing information regarding pre-existing knowledge on the upcoming lesson's subject matter or a one-minute essay on an index card summarizing their understanding of a key idea. A teacher may ask students to respond with physical cues (e.g., thumbs up/down/sidewise) or object cues (e.g., colored response cards) to indicate their level of understanding. A teacher may ask students to use electronic response devices to indicate their current understanding of a key concept. A teacher may ask for students to respond physically to a question before the class, perhaps by writing the answer to a problem on the board or acting out their perceived meaning of a verb. A teacher may use non-graded

quizzes during instruction to monitor student learning. In addition, the techniques may simply be those informal assessment conversations that occur naturally within the course of everyday classroom activity (Ruiz-Primo, 2011). The variety of techniques involving these pedagogical practices go beyond the scope of this study to fully explore; yet, the key point for this strategy of engineering effective classroom discussions, questions, and learning tasks is that these tools be used for the intentional purpose of monitoring student learning so that the learning gap may be closed (Pinchok et al., 2009). This is the heart of formative assessment.

Providing Feedback That Moves Learners Forward

The third operational scale of formative assessment is feedback. The strategy of *providing feedback that moves learners forward* focuses on the teacher's response to the monitoring of student learning. Feedback has been defined as "information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory, whether that information is domain knowledge, metacognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies" (P. H. Winne & Butler, 1994, p. 5740). Researchers have found that feedback is a key component in improving student achievement (Hattie, 2009); however, there are factors that affect its efficacy. Hattie and Timperley (2007), in examining multiple meta-analyses, including 196 studies and almost 7,000 effect sizes related to providing student feedback, concluded that the efficacy of feedback depended on factors such as the type provided (e.g., positive or negative) and the context in which it was provided (e.g., timing of feedback). Specifically, feedback is found to be most effective when it is specific, descriptive, immediate, and focused on student work rather than personal student characteristics (Chappuis & Stiggins, 2002).

Along those lines, Hattie and Timberley (2007) developed a model for effective feedback that incorporates three elements: learning goals, progress toward those goals, and steps needed to make better progress toward those goals. Interestingly, as seen in that model, the use of feedback is tied strongly to the first formative assessment strategy of delineating clear, mutually understood learning targets. Likewise, in outlining ten principles of learning, McTighe and Seif (2010) included models of excellence (per Strategy 1) and ongoing feedback (per Strategy 3) to enhance student learning, saying that "learners need to see models of excellent work and be provided with regular, timely, and user-friendly feedback in order to practice, retry, rethink, and revise their work" (p. 153). In describing the use of feedback, Hattie and Timperley (2007) gave specific examples of the types of feedback that will be effective. In contrast to giving feedback not explicitly tied to the task (e.g., *good job*), they recommended giving feedback regarding task performance (e.g., *incorrect*), task processing (e.g., *the problem would be easier if all fractions have the same denominator*), or self-regulation (e.g., *Can you think of a second method that will allow you to check your answers?*).

A recent study further supports the use of oral feedback to support student learning. A recent examination of 15 classroom-based studies (N =827) regarding the effectiveness of oral corrective feedback (CF) in second language acquisition (SLA) classrooms found that CF had significant and durable effects on target language development (Lyster & Saito, 2010). Types of CF included clarification requests, recasts, repetition, elicitation, metalinguistic clues, and explicit correction. In analyzing the results of that study, Lyster and Saito (2010) found that oral corrective feedback made a significant impact on second language learners (d = 0.74) for posttests in comparison

with control group students. Students who received CF displayed large effect sizes (d = 0.91) in comparison with their pretest performance. Students not receiving CF also exhibited improvement (d = 0.39), which may be attributed to test-retest effects or to the fact that these students, by virtue of being in classroom settings, also received instruction, albeit without intentionally designed CF treatments. While this study only considered the use of CF in second language classrooms, it does support the idea that oral feedback can result in enhanced student learning.

In reviewing the research on task-level feedback, Shute (2008) used the term *formative feedback* to describe "information communicated to the learner that is intended to modify his or her thinking or behavior for the purpose of improving learning" (p. 154). She described the complexity and variety of variables related to utilizing and evaluating use of feedback. Along those lines, Hattie and Timberly (2007) provided a summary of 74 meta-analyses studied done on feedback as related to those variables (see Table 2.1).

Table 2.1 Summary of effect sizes relating to feedback effects

Variable	Number of Meta-analyses	Number of Studies	Number of Effects	Effect size
Cues	3	89	129	1.10
Feedback	74	4,157	5,755	0.95
Reinforcement	1	19	19	0.94
Video or audio feedback	1	91	715	0.64
Computer-assisted Instructional feedback	4	161	129	0.52
Goals and feedback	8	640	121	0.46
Student evaluation feedback	3	100	61	0.42
Corrective feedback	25	1,149	1,040	0.37
Delayed versus immediate	5	178	83	0.34
Reward	3	223	508	0.31
Immediate versus delayed	8	398	167	0.24
Punishment	1	89	210	0.20
Praise	11	388	4,410	0.14
Programmed instruction	1	40	23	-0.04

Source. Hattie and Timberly (2007)

One variable, the issue of timing, seems particularly related to the use of an instrument to observe formative assessment in action. Is feedback more effective if it is immediate or if it is delayed? It appears that the answer is not consistently one way or the other. Shute (2008) provided an excellent summary of this issue:

> One way to resolve the inconsistency is by considering that immediate feedback may activate both positive and negative learning effects. For instance, the positive effects of immediate feedback can be seen as facilitating the decision or motivation to practice and providing the explicit association of outcomes to causes. The negative effects of immediate feedback may facilitate reliance on information that is not available during transfer and promote less careful or mindful behavior. If this supposition is true, the positive and negative effects of immediate feedback could cancel each other out. Alternatively, either the positive or negative effects may come to the fore, depending on the experimental context. A similar argument could be made for delayed feedback effects on learning. For example, on the positive side, delayed feedback may encourage learners' engagement in active cognitive and metacognitive processing, thus engendering a sense of autonomy (and perhaps improved self-efficacy). But on the negative side, delaying feedback for struggling and less motivated learners may prove to be frustrating and detrimental to their knowledge and skill acquisition. (p. 166)

For the purposes of this study, immediate feedback will be the focus due to the nature of the instrument as observational within a single instructional session. Potential issues associated with immediate feedback (e.g., lack of opportunity for metacognitive

processing), will be addressed through other strategies, primarily the fourth strategy of activating students as owners of their own learning.

As with communicating learning targets and monitoring student learning, various resources provide example of this enacting this strategy of providing feedback. One example of a formative assessment technique that supports this strategy is comment-only marking that provides non-graded feedback on assignments (Thompson & Wiliam, 2008). For some students, receiving this written feedback and being given the opportunity to reflect on it provides sufficient feedback. Others, however, may need face-to-face teacher feedback to reinforce what they have done well (Chappuis & Stiggins, 2002). Regarding the use of feedback, Linda Darling-Hammond (2010) stated:

> Effective assessment means assigning a piece of student work – whether it is an essay, a research project, a scientific inquiry, or a sculpture – and allowing a student to work on that selected task with support while scaffolding instruction and giving feedback that expands the student's understanding and skill. Teachers may combine peer assessment, student self-assessment, or their own assessment so that the students learn how to look at their work, learn strategies for framing and solving problems, and then understand how to continually revise their work so that they are getting closer and closer approximations to expert practice. (p. 39)

Activating Students as the Owners of Their Own Learning

The fourth operational scale of formative assessment is self-assessment. This strategy of *activating students as the owners of their own learning* focuses on developing students' self-regulatory abilities. The strategy of self-assessment seeks to encourage self-regulated learning (SRL), which has been described as academically effective forms

of learning that involve metacognition, intrinsic motivation, and strategic action (Winne & Perry, 2000). Research on self-regulated learning suggests that learning improves when teachers direct students to monitor their learning and show them how to achieve their learning objectives (Butler & Winne, 1995; Hyeon Woo, Kyu Yon, & Grabowski, 2010; Schunk, 1996). Accordingly, this self-assessment strategy asks students to use assessment tasks to answer questions such as (Chappuis & Chappuis, 2008):

- What are my strengths relative to the standards?
- What have I seen myself improve at?
- Where are my areas of weakness?
- Where didn't I perform as desired, and how might I make those answers better?
- What do these results mean for the next steps in my learning, and how should I prepare for that improvement?

This strategy recognizes the importance of students gauging their own growth over time, which can enable them to feel in charge of their own success and lay a foundation for life-long learning (Nicol & Macfarlane-Dick, 2006; Stiggins, 2002). Fortunately, teachers have reported that students' self-assessments are generally accurate, and students express that assessing their own work has helped them understand the material in a new way (Leahy et al., 2005).

Student self-assessment practices can be supported through the development of metacognitive skills in students (Dunlosky & Metcalfe, 2009; Flavell, 1979; Winne & Hadwin, 1998; Zimmerman & Schunk, 2001). Metacognition, or thinking about thinking, involves the knowledge of cognitive processes and products and the ability to

control, monitor, and evaluate those cognitive processes (Flavell, 1979). Winne and Hadwin (1998) described this metacognitive process as occurring in four stages: task definition (i.e., students determine their task for studying), goal setting and planning, enactment (i.e., the use of strategies to accomplish those plans), and adaptation (i.e., changes made to their learning process based on their experiences). The development of the metacogntive skills necessary to successfully accomplish these tasks can be a support for student learning, as a study utilizing the IMPROVE model of metacognitive training for students demonstrated. Mevarech and Fridkin (2006) studied 81 students involved in a pre-college mathematics course. The treatment group was trained to activate metacognitive skills through the IMPROVE model:

I – Introducing the new concepts

M – Metacognitive questioning

P – Practicing

R – Reviewing

O – Obtaining mastery

V – Verification

E – Enrichment and remedial

The control group received traditional instruction in the same problems with the same materials for the same amount of time (12 hours/week for a month). Both groups were given the same post-test at the end of the instructional period. The study found that, although there were no significant differences found between the two groups on the achievement mean scores prior to the beginning of the study [($F(1,79) < 1.00$, $p > .05$)],

an ANCOVA analysis of post-test results indicated significant differences between conditions at the end of the study, controlling for initial differences between conditions [($F(1,78) = 5.21$, $p < .05$)]. The study then attempted to further examine the effects of IMPROVE on mathematical knowledge and mathematical reasoning. One-way ANOVA and ANCOVA analyses indicated that although no significant differences between conditions were found prior to the beginning of the study [both $F(1,79) < 1.00$; $p > .05$)], significant differences were found at the end of the study [$F(1,78) = 10.14$; $p = .002$ on mathematical knowledge and $F(1,78) = 15.45$; $p = .001$ on mathematical reasoning]. The development of such metacognitive skills are important for successful student self-assessment and may provide an important tool for student learning.

Teachers can support the development of these abilities in students, and students need that support (Schunk, 1996). For example, through the use of observations and interviews in a qualitative study on self-regulated learning, Perry (2002) found children in kindergarten through Grade 3 can engage in self-regulatory behaviors, such as planning, monitoring, problem-solving, and evaluating, during complex reading and writing tasks. In investigating how teachers can best foster such behaviors, the study concluded:

> [Y]oung children can and do engage in SRL in classrooms where they have opportunities to engage in complex open-ended activities, make choices that have an impact on their learning, control challenge, and evaluate themselves and others. In addition, our observations revealed the ways in which teachers provide instrumental support to students (e.g., through questioning, clarifying, correcting, elaborating, modeling) and create opportunities for students to support one

> another (e.g., through collaborating, sharing ideas, and brainstorming problem-solving strategies). Last but not least, we observed how teachers created nonthreatening and intrinsically motivating learning contexts by embedding assessment and evaluation in the ongoing activities of their classrooms, making students accountable without being punitive, and encouraging students to focus on personal progress and view errors as opportunities to learn. (Perry, 2002, p. 14)

This exemplifies how students can take ownership of their learning and how teachers play a crucial role in that accomplishment. Earl (2013) wrote that these are "complex and difficult skills that do not develop quickly or spontaneously…becoming metacognitively aware requires modeling and teaching on the part of the teacher, and practice on the part of the student" (p. 53).

The development by students of the ability to self-regulate their learning is so important that Crisp (2012) has recommended creating a new term and category of assessment for it: *integrative assessment*. Specifically, he proposed the term integrative assessment to describe tasks whose primary purpose is to strengthen the ways students approach future learning by "providing activities that define and track strategies that students use to assess their own learning abilities and problem-solving capabilities, the quality and standards of student responses and how students might adapt their learning to future scenarios" (Crisp, 2012, p. 41). No matter the name attached, be it self-assessment or integrative assessment, there is widespread agreement regarding the importance of students' ability to monitor and respond to their learning progress.

Examples of formative assessment techniques that support this self-assessment strategy include students marking their work in specified ways to indicate their level of

understanding (e.g., drawing a smiling or frowning face on their work) and students writing daily learning logs summarizing their state of learning at the end of a lesson (Thompson & Wiliam 2008). Marzano (2006) suggested two methods for encouraging self-reflection by students. One is providing students opportunities to assign their own scores on assessments in relation to a grading scale provided to them. A second method is student articulation of their perceptions regarding their learning. This method may be accomplished in a number of ways, such as students writing a minute paper on their "muddiest point" of confusion, which the teacher would then use for further instruction.

It is important to note that connections exist between the role of self-assessment and other formative assessment strategies. For example, self-assessment connects to the first strategy of establishing clear learning targets. It has been observed that the main problem with student self-assessment is that they do not have a clear picture of the targets their learning is meant to attain (Black & Wiliam, 1998). Self-assessment also connects to the third strategy of feedback. Sadler (1989) made the case that feedback and self-monitoring are related to one another.

> For purposes of discussion, it is convenient to make a distinction between feedback and *self-monitoring* according to the source of the evaluative information. If the learner generates the relevant information, the procedure is part of self-monitoring. If the source of information is external to the learner, it is associated with feedback. In both cases, it is assumed that there has to be some closure of the gap for feedback and self-monitoring to be labeled as such. Formative assessment includes both feedback and self-monitoring. The goal of

> many instructional systems is to facilitate the transition from feedback to self-monitoring. (p. 122)

In a sense, everything that is done in connection to formative assessment has increased student ability to self-assess as its ultimate goal. As Bennett (2011) wrote, "sharing expectations, questioning, feedback, self-assessment, and peer assessment are intended to, among other things, help students develop internal standards for their work, reflect upon it, and take ownership of learning" (p. 9).

Activating Students as Instructional Resources for One Another

The fifth operational scale of formative assessment is peer-assessment. This strategy of *activating students as instructional resources for one another* focuses on the role that students can play in one another's learning. That role is connected to other formative assessment scales, especially strategies of self-assessment and establishing clear learning targets. Dylan Wiliam (2004) observed that learners often find it difficult to understand the criteria for success that the teacher has in mind; therefore, the involvement of peers can help learners understand success and monitor their own progress toward their goals. He proposed that peer-assessment not only provides a complement to self-assessment, but may actually be a prerequisite for effective self-assessment.

Peer learning can bring shown positive results, especially when thought is given to the issues such as context, objectives, curricular area, participants, helping techniques, length of contact, and resources needed (Topping, 2005). One example of the effectiveness of activating students as instructional resources for one another is found with a study done by Rust et al. (2003). He found in two cohorts that college students

who engaged in a peer process designed to increase understanding of grading criteria significantly increased achievement ($p < 0.01$) with an effect of .6 (cohort 1) and .69 (cohort 2). A key conclusion from that study was that "socialization processes are necessary for tacit knowledge transfer to occur" (Rust et al., 2003, p. 162), highlighting the importance of peer assessment for arriving at an explicit understanding of learning targets. It has been noted that "students from kindergarten to 12^{th} grade are much better at spotting errors in other students' work than in their own work," and thus, "peer assessment and feedback can be an important part of effective instruction" (Leahy et al., 2005, p. 23). A perhaps unexpected ancillary benefit of the use of peer assessment may be the enhancement of a student's abilities in self-regulated learning.

Teachers have available a variety of methods for enacting this strategy of peer-assessment in the classroom. Methods for initiating peer assessment of each others' work could include utilizing a preflight checklist of required components or utilizing a rubric describing the quality of those components (Thompson & Wiliam 2008). Another example could be the use of a homework helpboard on which students, when entering the classroom, write homework questions with which they struggled. Students then identify solutions and strategies for one another on that homework helpboard with minimal involvement by the teacher. Topping (2009) pointed out that peer assessment can vary across different curriculum areas, different outputs (e.g., writing, portfolios, oral presentations, and test performance), and different objectives (e.g., cognitive gains, metacognitive gains, or time savings). Whatever the method utilized, Russell et al. (2012) provide helpful guidance for enacting peer assessment in the classroom. They recommend that students be guided to focus on only one or two issues when assessing

each other's work. They also recommend that students, instead of making summative judgments of one another, be encouraged to identify effective elements in each other's work, point out places of confusion, and ask for the reasoning behind each other's decisions.

The strategy of peer assessment is not without challengers. Volante and Becket (2011) interviewed 20 elementary and secondary school teachers in two school districts in southern Ontario regarding their understanding and use of formative assessment strategies. While the study reported discomfort among many of the teachers regarding their ability to utilize self-assessment, it also reported that the consensus was that involving students in the assessment process is vital to student learning. On the other hand, peer-assessment was viewed much more problematically. The teachers noted their difficulties in the practical use of peer-assessment, including students' unfamiliarity with content material and students' lack of objectivity in giving feedback to one another. Nevertheless, while there may be challenges posed to the use of peer assessment in some classroom settings, the potential for gain remains.

Part 3: Existing Observational Instruments Related to Formative Assessment

Is formative assessment observable in practice? This is a crucial question because if we are to ascertain whether formative assessment truly makes a significant difference in student learning, we must have a method to identify its use.

Do such instruments exist? Do observational instruments exist that identify formative assessment use? In this part of the review of literature, I will examine existing observational instruments that include evaluations of formative assessment in classroom instruction. The nature and purpose of these instruments will be examined, as well as the

format for their implementation. As we will see, these instruments reveal that, while formative assessment forms an important and integral part of multiple existing instruments for evaluating teachers, there remains a need for an observational instrument designed specifically to identify formative assessment in practice.

In 2009, the Bill & Melinda Gates Foundation launched the Measures of Effective Teaching (MET) project. The purpose of the MET project was to improve the quality of information about teaching effectiveness by developing and testing multiple measures of teacher effectiveness ("Classroom observations," 2010). The MET project collected data across five research areas:

1. Student achievement gains on state standardized assessments and supplemental assessments designed to measure higher-order conceptual thinking
2. Classroom observations and teacher reflections
3. Teachers' pedagogical content knowledge
4. Student perceptions of the classroom instructional environment
5. Teachers' perceptions of working conditions and instructional support at their schools

The second of these research areas, classroom observations and teacher reflections, is closely related to the inquiry of this paper; as such, it provides particular assistance in determining relevant observational programs for examination.

The MET project enlisted the Educational Testing Service (ETS) to train and manage expert raters for observations of video-taped classroom lessons. The raters utilized two general observation protocols: Danielson's Framework for Teaching and the

Classroom Assessment Scoring System (CLASS). The raters also used content-specific observation protocols, including the Mathematical Quality of Instruction (MQI), the Protocol for Language Arts Teaching Observations (PLATO), and the UTeach Teacher Observation Protocol (UTOP) ("Gathering feedback," 2012). I will discuss the nature and purpose of each of these instruments, as well as the format for their implementation. In addition, I will discuss one other observational tool, developed by the World-Class Instructional Design and Assessment (WIDA) Consortium, which includes an examination of formative assessment.

The Framework for Teaching (FFT)

The Framework for Teaching (FFT) is a research-based protocol for evaluating teachers developed by Charlotte Danielson (Danielson, 2007) that has had widespread use. For example, Danielson's FFT has been adopted by the State of Idaho as the statewide foundation for teacher evaluation ("Teacher performance evaluation," 2012). Additionally, the FFT is aligned with the Interstate New Teachers Assessment and Support Consortium (INTASC) standards ("Danielson's framework," 2010). The FFT divides teaching into 22 components within four domains of teaching responsibility:

1. Planning and preparation
2. Classroom environment
3. Instruction, and
4. Professional responsibilities.

The MET project only used domain 2 (classroom environment) and domain 3 (instruction) in its observational evaluations. Of those two domains, domain 3 (instruction) is clearly most related to formative assessment. Within that third domain of

instruction, five key components are identified, along with several elements that comprise each component:

3a) Communicating with Students

- Expectations for learning
- Directions and procedures
- Explanations of content
- Use of oral and written language

3b) Using Questioning and Discussion Techniques

- Quality of questions
- Discussion of techniques
- Student participation

3c) Engaging Students in Learning

- Activities and assignments
- Grouping of students
- Instructional materials and resources
- Structure and pacing

3d) Using Assessment in Instruction

- Assessment criteria
- Monitoring of student learning
- Feedback to students
- Student self-assessment and monitoring of progress

3e) Demonstrating Flexibility and Responsiveness

- Lesson adjustment

All five elements contain, to varying degrees, aspects of formative assessment. For example, in Danielson's (2011) *The Framework for Teaching Evaluation Instrument*, element 3a (Communicating with Students) stresses the importance of clearly communicating learning goals to students and element 3e (Demonstrating Flexibility and Responsiveness) is described as a teacher's skill in making adjustments, which are clearly elements of teaching related to formative assessment. It is in element 3d (Using Assessment in Instruction) that Danielson makes the clearest allusion to formative assessment. She writes:

> Assessment of student learning plays an important role in instruction; no longer does it signal the end of instruction; it is now recognized to be an integral part of instruction. While assessment of learning has always been and will continue to be an important aspect of teaching (it's important for teachers to know whether students have learned what was intended), assessment for learning has increasingly come to play an important role in classroom practice. And in order to assess student learning for the purposes of instruction, teachers must have a "finger on the pulse" of a lesson, monitoring student understanding and, where appropriate, offering feedback to students. (Danielson, 2011, p. 62)

Raters using the FFT will typically utilize a three-step process of writing notes while observing, coding those notes for specific domains and components, and then rating the level of teacher performance for each component. Each lesson receives eight scores (one for each component). With the MET project, the scores from four such lessons are then combined.

The FFT, therefore, provides a broad perspective on teaching quality that clearly incorporates formative assessment use in that evaluation. However, providing one score for each component does not allow for the clear evaluation of specific formative assessment elements. For example, the FFT combines assessment criteria, monitoring of student learning, feedback to students, and student self-assessment and monitoring of progress into one single score. This does not allow for the depth of understanding and analysis of formative assessment use that an instrument focused on each of these operational elements might afford.

The Classroom Assessment Scoring System (CLASS)

Another general classroom observation tool selected by the MET project is the Classroom Assessment Scoring System (CLASS) ("Gathering feedback," 2012). CLASS is an observational tool that is based on research from the University of Virginia's Curry School of Education and has been studied in thousands of classrooms nationwide. The focus of the CLASS observation is explicitly on the daily interactions between students and teachers that are central to students' academic and social development. The data resulting from CLASS observations are intended for use in supporting teachers' unique professional development needs, setting school-wide goals, and shaping system-wide reform at the local, state, and national levels ("The CLASS™ tool," 2013).

The CLASS tool organizes teacher-student interactions into three broad domains that characterize students' experiences in school. Each domain includes several dimensions, some of which vary by grade level, that are defined by observable indicators ("The CLASS protocol," 2010).

The CLASS domains and dimensions are:

1. Domain 1: Emotional Support
    a. Pre-K and Lower Elementary – Positive Climate, Negative Climate, Teacher Sensitivity, Regard for Student Perspectives
    b. Upper Elementary and Secondary – Positive Climate, Negative Climate, Teacher Sensitivity, Regard for Adolescent Perspectives
2. Domain 2: Classroom Organization
    a. Pre-K and Lower Elementary – Behavior Management, Productivity, Instructional Learning Formats
    b. Upper Elementary and Secondary – Behavior Management, Productivity, Instructional Learning Formats
3. Domain 3: Instructional Support
    a. Pre-K and Lower Elementary – Concept Development, Quality of Feedback, Language Modeling
    b. Upper Elementary and Secondary – Content Understanding, Analysis and Problem Solving, Quality of Feedback, Instructional Dialogue

The third domain, Instructional Support, is most relevant to formative assessment, especially in its attention to the quality of feedback given.

Typically, the process for implementing the CLASS observational tool is as follows ("CLASS™," 2013):

1. Starting at the beginning of a school day, observe activity in the classroom for 20 uninterrupted minutes, paying special attention to the teacher's instructional interactions and behaviors; assign rating scores for each dimension on the Observation Sheet.
2. Repeat the observation-and-recording cycle up to six times during the school day for the most complete, accurate picture of teacher-student interactions.
3. Calculate scores across cycles and domains with the Scoring Summary Sheet for an at-a-glance look at areas of strength and weakness.
4. Use the results to inform program planning, shape in-service teacher training, and provide teachers with feedback that helps strengthen their skills.

Observers complete a two-day CLASS Observation Training that prepares observers to use the measure accurately. The training culminates with a test and one-year CLASS observer certification.

The MET project's utilization of the CLASS observation process called for observers to watch a video-taped lesson in 15 minute segments, scoring each segment with numerical codes for each of the CLASS dimensions, and averaging scores across the lesson. Four such lessons were scored for each teacher, and an average score for each dimension across the lessons was calculated. Scoring was done on a 7-point scale, with a low range being a score of 1-2, a middle range score being 3-5, and a high score being 6-7 ("The CLASS protocol," 2010).

Similar to the FFT, the CLASS observation process provides a broad perspective on teaching quality that includes aspects of formative assessment. However, the CLASS instrument seeks to accomplish a variety of purposes (professional development of teachers, school goal-setting, and broad-based educational reform) and to measure a variety of interactions within the classroom, including social development. While these are important aspects of the classroom learning experience and formative assessment use forms part of evaluating those aspects (as with the FFT), the CLASS instrument does not allow for the depth of understanding and analysis of formative assessment use that an instrument focused on each of these operational elements might afford.

The Mathematical Quality of Instruction (MQI)

The FFT and CLASS observational instruments are general tools designed to evaluate the totality of classroom instruction. The Mathematical Quality of Instruction (MQI) observational instrument, on the other hand, was specifically designed to measure mathematical work done in a classroom, and the MQI was selected by the MET project for the purpose of observing and evaluating mathematics instruction ("Gathering feedback," 2012).

The MQI was developed by Heather Hill and colleagues at the University of Michigan and Harvard University to provide scores for teachers on important dimensions of classroom mathematics instruction. They formed the MQI from the perspective that the mathematical work occurring in classrooms is distinct from classroom climate, pedagogical style, or the deployment of generic instructional strategies. For example, the presence of mathematical explanations and practices is scored separately from student

participation in mathematical explanations and practices ("Mathematical Quality of Instruction," 2012).

The MQI provides separate teacher scores for five important dimensions of classroom mathematics instruction, and it does so in the context of the relationships among the teacher, the students, and the content ("The MQI protocol," 2010). These five dimensions, with their designated relationships, include:

Teacher-Content Relationship

- *Richness of the Mathematics:*
    - Meaning-making includes explanations of mathematical ideas and drawing connections among different mathematical ideas (e.g., fractions and ratios) or different representations of the same idea (e.g., number line, counters, and number sentence).
    - Mathematical practices are represented by multiple solution methods, where more credit is given for comparisons of solution methods for ease or efficiency; by developing mathematical generalizations from examples; and by the fluent and precise use of mathematical language.
- *Errors and Imprecision:* Captures whether the teacher makes major errors that indicate gaps in his or her mathematical knowledge, whether the teacher distorts content through unclear articulation of concepts, and whether there is a lack of clarity in the presentation of content or the launch of tasks.

Teacher-Student Relationship:

- *Working with Students and Mathematics:* Captures whether the teacher accurately interprets and responds to students' mathematical ideas and whether the teacher corrects student errors thoroughly, with attention to the specific misunderstandings that led to the errors.

Student-Content Relationship:

- *Student Participation in Meaning-Making and Reasoning:* Captures the ways in which students engage with mathematical content, specifically:
    - Whether students ask questions and reason about mathematics; whether students provide mathematical explanations on their own or in response to the teacher's questions; and the cognitive requirements of a specific task, such as whether students are asked to find patterns, draw connections or explain and justify their conclusions.
- *Connections between Classroom Work and Mathematics:* Captures whether classroom work has a mathematical point, or whether the bulk of instructional time is spent on activities that do not develop mathematical ideas, such as cutting and pasting, or on non-productive uses of time, such as transitions or discipline.

Raters using the MQI divide each video-taped lesson into segments of approximately five to seven-and-a-half minutes and assign each segment with a score for each of the five elements, combining segment scores to create an overall score for the lesson. Scores of at least three lessons are averaged to yield a final teacher score ("The MQI protocol," 2010).

As can be seen from the elements above, formative assessment is embedded in elements of the MQI. For example, the element *Working with Students and Mathematics* clearly relates to monitoring and feedback and the element *Student Participation in Meaning-Making and Reasoning* relates to the strategy of self-assessment/metacognitive skills. However, the MQI does not focus on formative assessment per se. It intentionally focuses on a variety of elements related to mathematics instruction, such as a teacher's mathematical content knowledge. And, as previously mentioned, it takes the position that mathematical work is distinct from generic instructional strategies. As such, the need remains for a method of observing formative assessment.

The Protocol for Language Arts Teaching Observation (PLATO)

Just as the MET project utilized the MQI observational protocol for evaluating mathematics instruction, the project utilized the Protocol for Language Arts Teaching Observations (PLATO) for evaluating English Language Arts (ELA) instruction ("Gathering feedback," 2012). Pam Grossman, Professor of English Education at Stanford University, led the team that developed the PLATO protocol, as part of a research study on classroom practices in middle and high school ELA classes. The PLATO protocol scores elements of ELA instruction on a scale from one to four, with each element having been crafted to be as independent as possible from the others in order to capture different and independent aspects of classroom instruction ("Plato: Protocol," 2009). The elements are:

- Purpose focuses on the expressed clarity of ELA objectives, both in the short and long term;

• Intellectual Challenge focuses on the intellectual rigor of the activities and assignments in which students engage;

•Representation of Content captures the effectiveness of the teacher's explanations and examples in addition to his or her content knowledge;

•Connections to Prior Knowledge measures the extent to which new material is connected to students' previous academic knowledge;

• Connections to Personal and Cultural Experience focuses on the extent to which new material is connected to students' personal and cultural experiences;

• Models captures the availability of exemplars to guide student work;

• Explicit Strategy Instruction measures the teacher's ability to teach ELA strategies that can be used flexibly and independently;

• Guided Practice forces on the opportunities provided for students to practice ELA skills, concepts, or strategies in a structured and scaffolded way;

• Classroom Discourse reflects the opportunity for and quality of student conversations with the teacher and among peers;

• Text-Based Instruction focuses on how grounded ELA instruction is in a variety of texts.

• Behavior Management focuses on the degree to which behavior management facilitates academic work;

• Time Management focuses on how well-paced and efficient tasks and transitions are in the classroom.

The MET project, which utilized only eight PLATO elements (intellectual challenge, modeling, strategy use and instruction, guided practice, classroom discourse,

text-based instruction, behavior management, and time management), observed four video-taped lessons for each teacher. Each lesson was observed in multiple 15-minute independent observation cycles, with each element being scored for each cycle. Scores from each cycle and from each lesson (from non-consecutive days) were compiled to form the teacher's final score ("The PLATO protocol," 2010).

As with the MQI, the PLATO protocol includes aspects of formative assessment in its observational elements. For example, the element of Guided Practice relates to the formative assessment strategies of monitoring student learning and providing feedback, and the element of Strategy Use and Instruction relates to the strategies of self-assessment. However, as with the MQI's focus on mathematics instruction, PLATO's focus on ELA instruction means that it observes a breadth of classroom practices unrelated to formative assessment, such as Text-Based Instruction and Behavior Management. As such, the need remains for a method of observing formative assessment.

The UTeach Teacher Observation Protocol (UTOP)

In addition to mathematics and ELA, the MET project selected an observational protocol focused on instruction in science and math, the UTeach Teacher Observation Protocol (UTOP). The UTOP protocol was developed by the UTeach teacher preparation program at the University of Texas and was designed to value different modes of instruction, from inquiry-based to direct, in all age groups from K-college. It is structured in four ratings sections and uses a five point scale for rating different aspects of instruction within those sections ("Gathering feedback," 2012).

The four ratings sections of the UTOP protocol are Classroom Environment, Lesson Structure and Organization, Implementation, and Mathematics/Science Content. Within each section, specific indicators of success within that element are listed for observers to rate. For example, under Classroom Environment, one indicator is *The majority of students were on task throughout the class*. The UTOP includes eight indicators for Classroom Environment, six for Lesson Structure, nine for Implementation, and eight for Mathematics/Science Content ("The UTeach observation," n.d.). Of those sections and section indicators, a number of them address areas of formative assessment. For example, under section 2 (Lesson Structure), indicators included *the structure of the lesson included opportunities for the instructor to gauge*, and under section 3 (Implementation), indicators include *the teacher used formative assessment effectively to be aware of the progress of all students*, and *the lesson was modified as needed because the teacher was able to "read" the students' level of understanding through probing questions or other assessments of student understanding*. However, the much larger majority of indicators involve other areas of instruction, such as *the structure of the lesson uncovered important concepts in mathematics or science* and *the teacher had a confident demeanor* ("The UTeach observation," n.d.).

The UTOP asks observers to rate the indicators on a 5-point Likert Scale (1-5), with additional DK (Don't Know) and NA (Not Applicable) options. The scores are to be assigned after the observation has taken place and the observer has had an opportunity to review the video tape of the lesson and field notes. The numerical values for the Likert scale on the UTOP can be interpreted as follows:

1= Not observed at all/ Not demonstrated at all

2= Observed rarely/ Demonstrated poorly

3= Observed an adequate amount/ Demonstrated adequately

4= Observed often/ Demonstrated well

5= Observed to a great extent/ Demonstrated to a great extent

As it can be seen, each numerical value corresponds to two descriptors, one descriptor that measures the frequency of the occurrence of the indicator (observed rarely, observed often, etc.), and one descriptor that is intended to capture the quality of the implementation of that indicator (demonstrated poorly, demonstrated well, etc.). In addition to these rating, the UTOP observational report includes a post-observation teacher interview ("The UTeach observation," n.d.).

As with the other observational tools examined, UTOP does look for formative assessment in classroom instruction. However, while formative assessment strategies can be found within the UTOP observational tool, they are embedded within a wealth of other instructional areas (e.g., classroom management, lesson planning, and content knowledge). As such, there remains a need for a method to identify formative assessment use as a singular factor in instruction.

The World-Class Instructional Design and Assessment (WIDA) Consortium

In addition to the observational instruments chosen for inclusion in the MET project, I have included one additional instrument, the "Formative Assessment Best Practices Worksheet." This observational instrument was developed by the University of Wisconsin on behalf of the WIDA (World-Class Instructional Design and Assessment) Consortium ("Formative assessment best," 2009). WIDA is a respected resourcer of K-

12 education that seeks to advance academic language development and academic achievement for linguistically diverse students through high quality standards, assessments, research, and professional development for educators ("WIDA: World-class instructional," 2011). As part of their overarching mission, they developed an instrument for measuring formative assessment.

WIDA's observational tool frames formative assessment in a four part iterative cycle of goals, instruction, measuring, and feedback:

> First are instruction GOALS. These goals are based on relevant language learning targets, objectives or standards. It is best when these goals are shared by both teachers and students. Next is INSTRUCTION. Instruction is based on the pre-set learning goals and objectives. MEASURING is the third part of the assessment cycle. Measuring refers to the collecting of information about student learning. Are students meeting instructional goals? Are the instruments that are used to measure student language proficiency sufficient? The last part of the assessment cycle is FEEDBACK. This is a very important part of the cycle and often overlooked. What kind of feedback is provided to students? The goal of providing feedback is to promote action, action to set new goals or action to re-teach or re-instruct students to make sure they meet goals. ("Formative assessment best," 2009)

Clearly this cycle includes key strategies of formative assessment , including learning targets, monitoring student learning, and providing feedback. It does, however, omit the formative strategies of self-assessment and peer-assessment.

This worksheet includes a checklist of nine best practices, subdivided into elements to be rated as either *no*, *some*, *mostly*, or *yes*. These practices and their elements are as follows ("Formative assessment best," 2009):

I. Technically Sound

A. Valid – measure important concepts

1. Connected to meaningful learning targets & standards

2. Aligned to instructional goals

3. Focused on student learning needs

4. Appropriate measures of student performance

B. Reliable – provides consistent information

1. Item quality has been examined

2. Information from assessment provides actionable results for teachers & students

II. Embedded & Ongoing

A. Connected with curriculum

1. Part of the instructional process, not district from it

2. Connected to lesson plans, learning goals, and meaningful standards

B. Not “one-time-wonders”

1. Designed to be ongoing, iterative

2. A process, not just an event

III. Learning Goals

A. Connected to learning goals and targets

1. Aligned to standards and curriculum

2. Focused on student learning

3. Clear & explicit in what is assessed

4. Supports instructional goals

B. Organized to appropriate learning progressions

1. Based on appropriately sequenced language functions, vocabulary and/or grammar

2. Appropriate measures of students' current language learning goals

IV. Examples

A. For teachers & students

1. Rubrics, checklists, and rating scales have examples of each type of performance

2. Examples of "good student performance" are provided

V. Highlights Current Skills

A. Current Skills

1. Identifies with sufficient clarity, students' current abilities & skills: vocabulary knowledge, grammatical control, comprehension skills, communication skills, or discourse capabilities

VI. Highlights Future Goals

A. Future Goals

1. Identifies with sufficient clarity, students' future language abilities & skills: vocabulary knowledge, grammatical control, comprehension skills, communication skills, or discourse capabilities

2. Highlights next steps for students

VII. Integrated

A. Associated with other assessments used at the school, district and state

VII. Dynamic

A. Fits well into classroom realities (e.g., scheduling, timing)

B. Easy to administer & score

IX. Rigorous PD

A. Instrumentation development provided with adequate support

B. Structure in place to work with colleagues or professional learning communities in instrument development and scoring

As can be seen from this list of formative assessment practices and their elements, WIDA's observational worksheet intentionally focuses attention on formative assessment, and it includes positive, helpful elements, especially in the area of learning targets. However, it does omit key strategies of formative assessment, such as self-assessment and peer-assessment. Even more significantly, it appears to be oriented towards a curricular instrument-based perspective on formative assessment rather than an ongoing minute-to-minute perspective. For example, the worksheet asks whether measures are valid and reliable in terms of item quality, and it asks whether formative

assessments are easy to administer and score. These are questions appropriate for formal planned formative assessments rather than for the dynamic formative assessment that is an integral part of ongoing classroom instruction.

The strength of this tool is in supporting professional development and teacher planning, especially in the area of language learning goals and progressions. It is not well-suited for an inquiry into the use of formative assessment in ongoing classroom instruction or for research into formative assessment use. For example, for the tool to be maximally effective for research purposes, there would need to be clear descriptors of the ratings for each element to be rated. Consequently, it seems clear that there remains a need for a further work in forming a tool for observing formative assessment for research purposes.

Summary

Is formative assessment observable in practice? A multitude of programs exist for observing and evaluating classroom instruction. The question is whether those programs are able to effectively observe formative assessment in classroom education so that formative assessment's distinctive impact on student learning can be evaluated. Seeking an answer to that question, I have examined six different existing observational instruments. I chose five of the instruments based on their selection by the MET project for inclusion in their inquiry on measuring the quality of classroom instruction. In addition, I chose a sixth observational tool developed by WIDA that specifically focuses on formative assessment.

The MET project examined five classroom observation instruments for study: the Framework for Teaching (FFT), developed by the Danielson Group; Classroom

Assessment Scoring System (CLASS), developed by faculty at the University of Virginia; the Protocol for Language Arts Teacher Observations (PLATO), developed by Pam Grossman at Stanford; the Mathematical Quality of Instruction (MQI), developed by Heather Hill at Harvard; and UTeach Teacher Observation Protocol (UTOP), developed by faculty at the University of Texas-Austin. After employing and evaluating these instruments, the MET project reported its findings in January 2012 regarding the use of classroom observations to evaluate classroom instruction ("Gathering feedback for," 2012). The study found:

1. All five instruments were positively associated with student achievement gains.
2. Reliability characterizing a teacher's practice required averaging scores over multiple observations.
3. Combining observation scores with evidence of student achievement gains on state tests and student feedback improved predictive power and reliability.
4. Combining observation scores, student feedback, and student achievement gains was better than graduate degrees or years of teaching experience at predicting a teacher's student achievement gains with another group of students on the state tests.
5. Combining observation scores, student feedback, and student achievement gains on state tests also was better than graduate degrees or years of teaching experience in identifying teachers whose student performed well on other measures.

The report also emphasized three key take-aways: First, that high-quality observation will require clear standards, certified raters, and multiple observations per

teacher. Second, combining classroom observations, student feedback, and value-added student achievement gains capitalizes on teachers' strengths and offsets weaknesses. Third, combining new approaches to measuring effective teaching significantly outperforms traditional measures; therefore, providing better evidence should lead to better decisions ("Gathering feedback," 2012).

Results from the MET project also indicated that there is great room for expansion in the types of classroom practices in which formative assessment is used. In each of the observational instruments tested in the project, teacher practices were found to be strongest in areas such as managing student behavior and keeping students engaged. Teaching practices were weakest in areas such as the use of questioning/discussion, analysis/problem solving, strategy use, and feedback ("Gathering feedback," 2012). These findings highlight the need for increased formative assessment use that seeks to understand, support, and deepen student learning in such areas.

The results of my investigation into instruments currently being used in teacher evaluation demonstrate that formative assessment forms a crucial component in these instruments. However, the key is that it only forms a component. While formative assessment remains an integral part of current teacher education and evaluation, it is embedded in these instruments rather than standing as a factor to be evaluated on its own. Therefore, although evaluations of formative assessment by teachers are present within a number of programs aimed at evaluating and improving the quality of classroom instruction, there remains a need for an instrument with a specific purpose of observing the elements of formative assessment in action.

CHAPTER THREE: METHODOLOGY

Introduction

Is formative assessment observable in practice? The question is important because, while formative assessment practices can frequently be found embedded in both teacher training and evaluation, there remains an absence of an instrument/method specifically designed to evaluate its use by observing it in practice. In light of the potential impact of formative assessment on student learning, it is important that such an instrument be constructed and tested. Consequently, I undertook to develop and appraise an instrument designed to observe formative assessment in practice, thereby answering the question of whether formative assessment is observable in practice.

The purpose of this chapter is to describe the process by which I have developed and appraised the observational instrument in terms of its validity and reliability. It is appropriate to separate the validation process into the two stages of development and appraisal, even though they may overlap at times (Kane, 2006). Consequently, the chapter will contain two parts: Part 1: Instrument Development and Part 2: Instrument Appraisal. In order to heighten study validity and make a clear chain of reasoning (Krathwohl, 1989), I will describe various components involved in this study, including components of formative assessment, instrument design, validity, reliability, participants, and collection of data.

Instrument Development

Variables

In this study, I attempted to determine whether it is possible to observe formative assessment in practice. In order to make that determination, I developed an observational instrument for observing formative assessment in practice. I attempted to establish the instrument's validity and reliability as I employed the instrument for in situ observations of elementary classroom teaching. Through this process, it was my goal to create an instrument that future researchers may use to evaluate the efficacy claims for formative assessment and to deepen understanding of various components of formative assessment. It was for this purpose that I based the variables to be considered, seeking to ensure that the variables of interest were both identified and operationalized in such a way that data collection yields useful information (Horn, Snyder, Coverdale, Louie, & Roberts, 2009). In order to accomplish that purpose, therefore, I identified the construct of formative assessment and operationalized it into observable components (Kimberlin & Winterstein, 2008). The foundational process of distinguishing the construct of formative assessment and operationalizing into observable components relied upon extant scholarly work done in the area of formative assessment, collaboration with experts in the field of education, and field testing of the instrument.

As previously discussed, this study relied heavily on the work of Dylan Wiliam (Wiliam, 2010) in both defining and operationalizing formative assessment. Measurement of a construct requires that its conceptual definition be translated into an operational definition (Kimberlin & Winterstein, 2008). Consequently, teacher use of formative assessment was evaluated on the basis of five components, which are the five

operationalized formative assessment strategies previously discussed and exemplified (Leahy et al., 2005; Wiliam, 2010):

1. Learning Targets: Clarifying and sharing learning intentions and criteria for success.
2. Monitoring Student Learning: Engineering effective classroom discussions, questions, and learning tasks.
3. Feedback: Providing feedback that moves learners forward.
4. Self-Assessment: Activating students as the owners of their own learning.
5. Peer-Assessment: Activating students as instructional resources for one another.

Using these five formative assessment strategies as the key components operationalizing formative assessment, I developed an observational instrument around them to be used in identifying the use of formative assessment. Within the observational instrument, I included observational items for each component that serve as indicators of that component's use.

Instrument Design

In this study, the tool for identifying the presence of formative assessment in action was an observational instrument. Assessment through structured observation is a legitimate and commonly-used technique in education, both for the evaluation of students and the evaluation of teachers (Danielson, 2012; Leff, Thomas, Shapiro, Paskewich, Wilson, Necowitz-Hoffman, & Jawad, 2011; Russell et al., 2012; Shapiro, 2004). Additionally, observational instruments have been demonstrated to be an effective strategy for improving teacher quality when used with feedback (Allen, Pianta, Gregory,

Mikami, & Lun, 2011). The purpose of the observational instrument I developed during this study was to observe formative assessment in practice.

I constructed this observational instrument around the five aforementioned operational components of classroom formative assessment (Learning Targets, Monitoring Student Learning, Feedback, Self-Assessment, and Peer-Assessment), and I developed observational items as indicators of each component's presence. Since there is no existing instrument that focuses solely on observing formative assessment, I looked for guidance from various sources in formulating these observational items. Primarily, I looked to extant research in formative assessment as presented in Chapter 2, Parts 1 and 2. I also looked to relevant resources on teacher evaluation, such as the Danielson Framework for Teaching (Danielson, 2007) which is commonly used for teacher evaluation ("Teacher Performance Evaluation," 2012). Additionally, I looked at the methods and outcomes of the MET project discussed in Chapter Two, part three. Upon this basis, I developed the list of observational items for each formative assessment component whereby teacher use of that component of formative assessment was measured. I designed these observational items so as to be answered solely through classroom observations, with each item operating distinctively within its attendant component.

After creating an initial set of potential observation items, I went through an iterative process of review and revision with professors of education at an urban state university in Mountain West region of the United States. Through that process, I reduced an initial list of potential items down to a group of 28, divided into five groups corresponding to the five formative assessment components. I endeavored to order those

observational items within each component in such a way that they progress from less advanced to more advanced usage of the formative assessment component, as determined by reviewing the literature on formative assessment use. Through this collaborative process of validation and field testing, I endeavored to narrow the number of items ultimately to a total of between 15 and 25, which would strike a balance between making the instrument short enough to be easily used in practice and yet long enough to measure the components reliably.

Validity

In the creation of observational instruments, validity refers to "the degree to which scores represent the underlying construct they seek to measure" (Hill, Charalambous, Blazar, McGinn, Kraft, Beisiegel, & Lynch, 2012, p. 89), and Kane (2006) wrote that the first step in validating any proposed interpretation of construct measurements is "to evaluate the coherence and completeness of the proposed interpretive argument" (p. 43). He went on to say:

> The interpretation of indicator scores as estimates of a theoretical construct extends the interpretation to a claim about a construct as defined by the theory. Theory-based interpretations of indicator scores assume that the theory provides a sound explanation for the relevant phenomena and that the indicators provide appropriate estimates of the constructs in the theory. The warrant for this inference is the theory. (Kane, 2006, p. 44)

Having accomplished this step through establishing the theoretical foundation for this instrument in Chapter Two, I undertook to establish validity for this instrument as warranted by that theory.

Three major methods of establishing validity are content validity, criterion-related validity, and construct validity (Allen & Yen, 1979), although it may be argued that both content and criterion-related validation strategies can be subsumed within an all-encompassing view of construct validity (Kane, 2006). For the purposes of validating an observational instrument to identify the presence of formative assessment in action, I relied primarily upon content validity. Cronbach (1960) described content validity as a legitimate means of establishing validity when the question is whether a test represents the content or activities intended to be measured. He wrote, "Instead of comparing scores on the test with some other measure or judgment, as in empirical validation, he must examine the items themselves and compare them with content he wished to include. This process is called *content validation*." (Cronbach, 1960, p. 104).

Allen and Yen (1979) wrote that there are two main types of content validity, face validity and logical validity, and that content validity is established "through a rational analysis of the content of a test, and its determination is based on individual, subjective judgment" (p. 95). The first type of content validity, therefore, is face validity, which is established when a person examines a test and concludes that it does measure the trait in question (Allen & Yen, 1979). Face validity, however, is not without its weaknesses. A layman in the field may mistakenly view a test as plausible and reasonable (Cronbach, 1960); therefore, in utilizing face validity in the creation of an evaluatory instrument, it is important that those providing face validity be professionals in the field of inquiry. Consequently, in order to avoid the potential pitfalls of face validity, I only included individuals who were knowledgeable and experienced in the field of Education to provide confirmation of validity.

A second type of content validity is logical or sampling validity, which involves "the careful definition of the domain of behaviors to be measured by a test and the logical design of items to cover all the important areas of this domain" (Allen & Yen, 1979, p. 96). Establishing validity through the careful definition and delineation of an instrument's domain, purpose, and scope corresponds with Harlen's (2007) comments regarding the validity of student assessment systems:

> The important requirement is that the assessment concerns all aspects – and only those aspects - of students' achievement relevant to a particular purpose. Including irrelevant aspects is as much a threat to validity as omitting relevant aspects. Thus a clear definition of the domain being assessed is required, as is adherence to it. (p. 18)

In Chapter Two, I provided a clear definition of the domain to be assessed, which is formative assessment. Adherence to that domain and to the specific components therein has been maintained throughout the validation process.

In the development of this instrument seeking to observe formative assessment in practice, I have attempted to establish validity through examining the relationship between the content of the instrument and the construct it is designed to measure (Reynolds, Livingston, & Willson, 2009). Evidence regarding that relationship was gathered through consulting experts in the field and enlisting them to review the instrument and demonstrate content-validity (i.e., that it actually measures the construct intended) (Reynolds et al., 2009). In order to gather evidence regarding content validity, I provided the suggested 28 items, grouped into five formative assessment components, to faculty members at the school of education of an urban state university in the

Mountain West of the United States for review. They reviewed and approved of the suggested potential items and their grouping, providing a basis for content validity for the instrument.

In order to further establish validity empirically for this instrument, I conducted a card sort exercise, which has been demonstrated to have the potential of adding to the validity of research (Faiks & Hyland, 2000; Jahrami, Marnoch, & Gray, 2009). The practice of sorting objects into groups has been commonly used in the cognitive and social sciences since the 1950s, and it has been defined as a method for "putting a number of things into a smaller number of groups and being able to give the rule by which such allocation is made" (Coxon, 1999, p. 1). A card sort exercise asks participants to impose their own categorical organization on a set of items and concepts. The exercise typically provides a group of participants with a set of cards. Written on each card is a concept or piece of information from the set that needs to be organized. The participants then sort the cards with similar concepts into piles. A card sort exercise is based on the assumption that if users (assuming they are knowledgeable in the field) group cards together, the concepts probably should be grouped together (Faiks & Hyland, 2000).

For this card sort exercise, I first conducted a preliminary card sort in which each of the initial 28 items were printed on individual cards. Those cards were shuffled and given to groups of two or three experienced educators who were enrolled in a doctoral program in education. These groups were asked to assign the cards to one of the five formative assessment components. Each group was given a separate set of cards and was encouraged to make notations on the cards regarding any points of confusion or lack of

clarity. Based on the results of this preliminary card sort, I eliminated five items and made further revisions, narrowing the total number of potential items to 23.

Using these 23 items, I then completed a second card sort. Again, each of the remaining observational items was printed on individual cards. Sets of those cards were shuffled and given to six pairs of professional educators, who were asked to assign the cards to one of the five formative assessment components. Of those six pairs, five categorized the cards with 100% accuracy and one was accurate on 22 of 23 cards, resulting in an overall agreement rate of 99%. These results support confidence in the content validity of the instrument.

Rating Scale

An important consideration in designing this instrument was the manner in which raters indicate formative assessment in practice. Specifically, the question is whether to ask raters to respond to the observational items with a dichotomous "yes/no" answer or with a rating on a 5-point Likert-type scale. Likert scales were developed by Rensis Likert (1932), and they typically provide a range of responses, frequently ranging from 1 to 5 with 1 signifying strong disagreement and 5 signifying strong agreement (Jamieson, 2004). Research regarding the use of Likert scales has produced seemingly contradictory finding regarding the optimal number of categories (e.g., 3-point, 5-point, 7-point, etc.) to be included in a scale (Croasmun & Ostrom, 2011; Guilford & Guilford, 1954; Matell & Jacoby, 1972; Ray, 1980). Cronbach (1960) appeared to favor the 5-point scales, stating that it "obtains more discrimination than the 'yes-no' checklist" (p. 511).

In a comparison of the two response strategies, Greenwald and O'Connell (1970) reported that dichotomous measures (e.g., true-false, yes-no, agree-disagree scales) yield

similar but not equivalent information to that of Likert scales. They also pointed out that each approach has disadvantages: "dichotomous approaches can force inadvertent responses, distort bona fide neutral responses and falsely generate extreme total scores, while Likert scales can heighten response variability, diminish stability and falsely imply precision" (Greenwald & O'Connell, 1970, p. 481). Thus, choice of response method may be best determined by considering the purpose for which the instrument is to be used.

The purpose of this instrument within the larger context of the project with which it was associated was to provide a basis for comparing the presence of formative assessment with teacher accuracy in predictions of student performance. A secondary purpose for this instrument in the future may be to provide feedback to teachers regarding their formative assessment practices. For these purposes, I have provided both a 5-point Likert scale that may provide guidance for future research and professional development in formative assessment, and I have provided a dichotomous yes/no subgrouping within that scale that may be utilized in conducting research in the relationship of formative assessment presence and other educational factors, such as teacher monitoring of student learning. Jacoby and Matell (1971) found that "investigators would be justified in scoring Likert-type scale items dichotomously (or trichotomously), according to direction of response, after they have been collected with an instrument that provides for the measurement of direction and several degrees of intensity" (p. 499). By creating a five point scale with a yes/no subgrouping within that scale (1-2 = no; 3= uncertain; 4-5 = yes), I hoped to also provide another method for establishing reliability.

I designated the midpoint rating of 3 as uncertain, thus to be disregarded in making the dichotomous yes/no determination. I have done this because a 3 rating on formative assessment use is so minimal that it may be difficult to distinguish from other areas of instructional practices. Not only is a 3 rating difficult to distinguish as a separate indicator, the use of formative assessment is often very minimal (even if somewhat present). And because it is so minimal, it likely has very minimal (perhaps indiscernible) influence. The ultimate educational question is whether effective use of formative assessment has a positive influence on student learning, and perhaps whether its complete absence has an adverse (or at least non-advantageous) effect. A 3 rating reflects such minimal or uncertain use that it may confuse the matter either way. By removing the '3's from the yes/no decision, I propose that clearer answers will emerge from future research into the effectiveness of formative assessment.

Consequently, each of the 3-5 observational items that comprise the use of a specific formative assessment component were rated on the basis of a dichotomous yes/no basis and 5-point Likert scale as follows:

1 = No evidence of use (No)

2 = Superficial or ineffective use (No)

3 = Minimal use or uncertain effectiveness

4 = Frequent or effective use (Yes)

5 = Pervasive or highly effective use (Yes)

Observer responses to these items regarding specific aspects of the given formative assessment components indicate whether, and to what degree, those components and their comprising items were used during instruction.

In order to facilitate accurate identification of a teacher's use of each formative assessment component, it was critical that there be clear descriptions of performance levels for each item. Consequently, I created an observational protocol that includes descriptors of each rating level (1-5) for each item. These descriptors included both quantitative (e.g., minimal, frequent) and qualitative (e.g., ineffective, effective) language where appropriate (Danielson, 2012). These descriptors were created on the basis of the theoretical research in formative assessment described in Chapter Two.

Reliability

Reliability refers to the consistency of results we obtain from an assessment. This may involve consistency across time, consistency across tasks, and consistency across raters (Darr, 2005). Consistency across time and consistency across raters were relevant to the development and appraisal of an observational instrument; therefore, I evaluated whether the results of the observation was similar for one person rating the same instructional sessions at different times and for more than one person observing the same instructional session. These are questions of rate-rerate reliability and of inter-rater reliability, or inter-observer agreement, which establishes the equivalence of ratings obtained with an instrument when used by different observers (Kimberlin & Winterstein, 2008).

In any method involving the use of judgment by observers, there is potential for error. These may include sources of error such as generosity errors (i.e., the tendency of raters to give favorable reports), ambiguity errors (i.e., unclear rating standards), constant errors (i.e., individual tendencies to rate high or low), and the halo effect (i.e., rating specific traits on the basis of a general opinion about the person's merit) (Cronbach,

1960). One way I attempted to address such potential problems was through intentionally utilizing raters who were knowledgeable in the field (Cronbach, 1960). Raters were faculty and graduate students from the College of Education at a public urban university in Mountain West of the United States. These raters were individuals who were involved in a federally funded project researching formative assessment entitled Improving Teacher Monitoring of Learning (ITML). Consequently, the raters involved in utilizing this instrument were considered knowledgeable in the field of formative assessment; thereby, reducing the potential for error.

I also attempted to address potential problems with observer ratings through carefully preparing a rating scale (Cronbach, 1960). Inter-rater reliability for an observational instrument relies on the development of precise operational definitions of the variables being measured and on having observers who are well trained in using the instrument (Kimberlin & Winterstein, 2008). In order to develop such precise operational definitions, I created an observational protocol for the use of the formative assessment observational instrument. That protocol provided a descriptor for each 1-5 Likert scale rating for each item. Thus, each item related to the five formative assessment components included 5 descriptors, one for each possible ranking. In order to have observers who were well trained in using the instrument, I provided this protocol to observers in advance and provide training (one-on-one or in a group or both) in the use of that protocol.

After training in the instrument and the protocol, inter-rater reliability among observers was established through a process of viewing in-person and video-taped classroom teaching. Using videos or in-person observations of elementary classroom

mathematics instruction, pairs of raters individually employed the instrument to evaluate formative assessment use. I then compared ratings across raters.

Field Testing

Field testing provides the opportunity to employ and evaluate the observational instrument in action in the context for which it is designed, which is real-life classrooms. The field testing of this instrument consisted of two parts. For the first part, a fellow doctoral student in the field of education and I conducted paired in-person observations of elementary classroom instruction. The process involved four parts: Observe-Rate-Compare-Revise. We jointly observed in situ instructional sessions. After each session, we individually used the instrument to rate the use of formative assessment during that instructional session. We then compared our ratings, discussing the reasons for any differences in rating (e.g., divergent expectations of teachers, unclear wording, unforeseen classroom practices, etc.) and noted points of needed clarification or revision in the instrument. I then revised the instrument based on what I learned through the cycle of observing, rating, and comparing. After making revisions, we repeated the process of observation, rating, comparing, and revising the instrument. In total, we field tested the instrument in eight in situ classroom sessions.

The second part of the field testing process involved the same Observe-Rate-Compare-Revise process. For this part, however, a member of the educational faculty and I independently viewed videos of elementary classroom instruction, rated them individually, and compared results. Based on those results and the subsequent discussions, I made further revisions to the instrument. As a result of this entire field testing process, three items were removed due to redundancy or unclarity and others were

revised to improve clarity and accuracy. Consequently, following the field testing process, the formative assessment observational instrument came to consist of 20 items divided into five formative assessment components. Also of note, key points discovered through both parts of the field testing process were incorporated into the training for those who were to be utilizing the observational instrument.

Instrument Appraisal

Rater Training

Prior to conducting observations, raters participated in internal training to ensure that raters interpret component item ratings and descriptors similarly. During a three hour instructional session, I facilitated training of potential raters in the instrument (see Appendix A) and an observational protocol providing guidelines, or descriptors, for assigning numerical ratings to each item (see Appendix B). The potential raters for utilizing this instrument included faculty and graduate students experienced in the field of education and knowledgeable regarding formative assessment. Consequently, I did not need to include time for training in formative assessment theory and practice in providing training for this group of observers. The training provided familiarity with the observational instrument and clarity regarding the observational items therein. We discussed in depth the observational protocol that describes the standards for ranking each item, and I answered questions regarding their delineation. During this training, I planned for team members to observe and rate videotapes of elementary classroom mathematics instruction and to compare their interpretations and ratings; however, time did not permit. Nevertheless, through our discussions we were able to arrive at a

common understanding of the theoretical framework, terminology, and rating levels. I made myself available for future consultation or additional training as needed to clarify and align interpretations of observation indicators and terminology to ensure continuing inter-rater reliability.

Participants

The design of the study attempted to involve as participants 23 teachers at four different elementary schools in a single district within a metropolitan region of a Mountain West state in the U.S during the 2012-13 school year. These schools and teachers were participating in the first year of a three year project entitled ITML. This project, funded by the Institute of Educational Sciences (IES), was investigating formative assessment and its relationship to the accuracy of teachers in predicting student achievement. The ITML project included 96 teachers distributed equally among eight different randomly chosen elementary schools in the same school district.

While originally this study was designed to observe 23 different teachers at four different schools, teacher schedules, rater availability, and limited resources reduced the number of teachers and schools involved in the study. Ultimately, sixteen teachers at three different schools were observed as part of this study. These teachers were chosen primarily on the basis of convenience as they were already participating as part of the larger federally-funded research project studying formative assessment use. The selection was also based on the teachers' willingness to be observed and/or videotaped. Their selection took into account a number of other factors such as school class schedules, teacher scheduling conflicts, and other time constraints. Ultimately, their

selection was based on their assignment to this study from the ITML leadership group, and as such, their selection was not up to this researcher's discretion.

Of the sixteen teachers observed as part of this study, fifteen were female and one was male. All teachers were Caucasian. The teachers worked at three different elementary schools in the same school district in a suburban metropolitan community of the Mountain Western part of the United States. Teachers were evenly distributed across the schools, with two schools having five teachers involved and one school having six teachers involved. Teachers in the study were also distributed across grade levels. Two teachers taught Kindergarten, three taught 1st Grade, two taught 2nd Grade, four taught 3rd Grade, two taught 4th Grade, and three taught 5th Grade. Of the 12 teachers who provided information regarding years of experience and highest level of formal education attained, three teachers had received Master's degrees in addition to their Bachelor's degrees that they all possessed. Those 12 teachers ranged from 6 to 31 years of experience, with an average of nearly 15 years of experience per teacher.

Data Collection

For this study, I developed an instrument with the goal of determining whether formative assessment is observable in practice. In order to gather data on the observability of formative assessment and on the reliability of the instrument in evaluating it, pairs of raters were assigned to utilize the instrument during classroom instruction. Raters conducted these observations during single mathematics instructional sessions in the course of regular class instruction. In other words, the instructional sessions observed were expected to be typical of the teacher's instruction and were not intended to disrupt the normal class routine. These observations took place over the

course of two months during the Spring semester of the 2013 school year. It should be noted that due to the structure of the ITML project, the observations were limited to class sessions involving mathematics instruction.

In order to better gain an accurate understanding of each teacher's use of formative assessment, each teacher was observed more than once. Fifteen of the teachers were observed three separate times and one teacher was observed twice, resulting in a total of 47 observational sessions. Most of these observations were approximately 30 minutes in length, although classroom schedules resulted in two observations only being 10-15 minutes in length. However, the observers decided that they had adequate information from those shortened lessons to make accurate evaluations.

Observations were conducted by a team of three graduate students who were involved in the aforementioned federally-funded research project on formative assessment. As such, each observer possessed a solid understanding of formative assessment and of the instrument to be utilized. For each of the 47 instructional sessions, two raters observed and independently rated the teacher's use of formative assessment as detailed in the observational instrument. In order to provide a consistent baseline for comparison, I observed and rated each of the 47 sessions myself. The other two raters shared the responsibility of providing the second rating for each session.

Due to logistical constraints of the study, a combination of real-time, in person observations of classroom instruction and later video-taped observations of classroom instruction were used. Occasionally, due to a teacher's unwillingness to be videotaped, both raters were present in the classroom at the same time (N=6). The vast majority of the time, however, the evaluations of formative assessment use was a combination of in-

person and videotaped observations (N=41). The normal procedure was for the assigned observer to use a digital video camera to record the instructional session. After the session, the observer completed the observational instrument immediately or could re-watch portions of the video if needed. The recorded lesson(s) were transferred to a portable hard drive for archiving and for sharing with a second rater. A second rater received a hard disk drive containing the recorded lesson, watched the lesson, and then completed the observational instrument. Whether conducting the observation live in the classroom or through video recordings, observers took detailed notes relating to the content of the lesson and the specific items on the observational instrument. Those notes were then used in helping the observers complete their ratings. After completing ratings of individual sessions, raters would e-mail the completed observational instrument to me. I then entered the results directly from those completed instruments into an Excel spreadsheet for analysis. The analysis included inter-rater reliability measures of exact agreement and Cohen's kappa, and the analysis also included internal consistency measures of Cronbach's alpha.

CHAPTER FOUR: RESULTS

Introduction

Is formative assessment observable in practice? If the answer is yes, then observers should be able to identify and evaluate its use in a classroom setting. If the answer is yes, then observers should be able to identify its presence and the degree to which it is utilized with reliability across time and across raters. If the answer is yes, then the instrument used to identify its use should have internal consistency regarding the areas it delineates as different components of formative assessment.

This chapter examines the results obtained from testing an observational instrument designed to identify formative assessment in practice. Ultimately, the reliability of this instrument in accomplishing that task will support an answer to the question of whether formative assessment is observable in practice. The hypothesis was that formative assessment is, in fact, observable in practice.

As discussed in Chapter Two, formative assessment can be operationalized into five components. This instrument was intended to capture data on those five major components of formative assessment use. These five components are titled:

1. Learning Targets: Clarifying Learning Intentions and Sharing Criteria for Success

2. Monitoring: Engineering Effective Classroom Discussions, Questions, and Learning Tasks that Elicit Evidence of Learning

3. Feedback: Providing Feedback that Moves Learners Forward

4. Self-Assessment: Activating Students as the Owners of Their Own Learning

5. Peer Assessment: Activating Students as Instructional Resources for One Another

Each of these five components forms a separate construct, thus each component will be measured with a separate scale. These five scales will be made up of 3-5 different indicators or items that are each measured on a Likert-type 1-5 scale (where 1 = no evidence of use and 5 = pervasive or highly effective use). As these five components (hereafter also referred to as scales) form five scales of measurement in this instrument, I examined data from the instrument as a whole and from each of these five scales. Where appropriate, I also included data on individual items.

Almost all assessments are based on samples. The sample may be answers on a mathematics test, the performance of a piece of music, or a session of observed classroom instruction. Based on a sample, an inference is made regarding the quality of whatever is being assessed, be that mathematical knowledge, musical skill, or use of formative assessment. In other words, assessment is an inference-based process. Since room for error already exists within the inferential nature of assessment, it becomes even more important that sources of error be minimized in the sampling process.

In the process of assessment, inaccuracy may enter in at various points. For example, inaccuracy may enter through the theoretical foundation of the assessment itself. Inaccuracy may enter through the design of the assessment. Inaccuracy may also enter through the way in which the assessment is utilized. In order to make the most

accurate inferences, the sampling process must be designed to reduce as many sources of potential error as possible.

In the attempt to evaluate teacher use of formative assessment, an observational approach to assessment was taken. The theoretical basis for the instrument and for what it attempted to assess was described in Chapter Two. The methodological process for designing the instrument was described in Chapter Three. The issue for this chapter is then to describe the way the instrument was used and what that reveals about the reliability of the instrument in providing an accurate sample from which to make inferences.

The primary question is one of evaluating possible measurement error. Such error can be an impediment to presenting an accurate rating of a subject and can be introduced in three ways. In an overview of computing inter-rater reliability for observational data, Hallgren (2012) summarized:

> Measurement error (E) prevents one from being able to observe a subject's true score directly, and may be introduced by several factors. For example, measurement error may be introduced by imprecision, inaccuracy, or poor scaling of the items within an instrument (i.e., issues of internal consistency); instability of the measuring instrument in measuring the same subject over time (i.e., issues of test-retest reliability); and instability of the measuring instrument when measurements are made between coders (i.e. issues of IRR). Each of these issues may adversely affect reliability… (p. 24)

In analyzing the data resulting from the use of this instrument, I attempted to address each of these potential sources of error.

Reliability Across Time

One source of potential error in the reliability of this instrument is the rater himself or herself. The question is how consistently the same rater can evaluate the same instructional session at two different times (cf. test-retest reliability). In other words, how much will a person's evaluation of formative assessment usage vary over time when using this instrument to observe the same lesson twice?

To evaluate rater re-rater reliability, I randomly selected half of the teachers used in the study, whose lessons were video recorded. I watched and re-rated one instructional session (the second of the three original observations) from each of those teachers. I recorded my responses and entered them into an Excel spreadsheet for comparison with my previous ratings of those same instructional sessions.

In order to determine the level of rater re-rater agreement, I compared the first and second ratings of the selected instructional sessions, ratings made using a Likert-type 1-5 scale (where 1 = no evidence of use and 5 = pervasive or highly effective use). I then computed the rater re-rater agreement between those rating by calculating the percentage of perfect agreement and agreement within 1 point for each item, each scale, and the instrument as a whole. In order to make that calculation, I divided the number of items receiving the same score by the total number of items, and I then multiplied the result by 100. In order to determine the level of agreement according to the less rigorous criterion of scores within one point of each other, I divided the number of items scored within one point of each other by the total number of items and then multiplied the result by 100 (Reynolds et al., 2009).

As another measure of rater re-rater reliability, I calculated Cohen's kappa for each scale and for the instrument as a whole (Cohen, 1960). Cohen's kappa is a commonly used statistic for assessing reliability for nominal categories, and it is used to correct for the amount of agreement between observers or observations that would be expected by chance (Hallgren, 2012). In addition to Cohen's kappa, I also made calculations of Cohen's weighted kappa (Cohen, 1968), which takes into account varying degrees of agreement or disagreement in nominal scale assignments. It is appropriate to include weighted kappa in evaluating rater reliability with a 5-point Likert-type scale because:

> In case categories are ordered along a continuum of values, it is desirable to give partial credit for near agreement. Because weighted kappa allows for differential weighting of disagreement, it is an attractive agreement statistic for ordered categories and preferable to Cohen's kappa, which distinguishes only between agreement and disagreement cases. (Schuster, 2004)

The possible values for kappa statistics can range from -1 to 1, with -1 representing perfectly consistent disagreement, 0 representing completely random agreement, and +1 representing perfectly consistent agreement. A guideline for the interpretation of kappa values has been provided by Landis and Koch (1977), with kappa values from 0.0 to 0.2 representing slight agreement, 0.21 to 0.40 representing fair agreement, 0.41 to 0.60 representing moderate agreement, 0.61 to 0.80 representing substantial agreement, and 0.81 to 1.0 representing almost perfect or perfect agreement.

In addition to scoring formative assessment use with a 5 point Likert-type scale, I also created a dichotomous scoring model using a yes/no subgrouping within that scale

(where 1-2 = no; 3= uncertain; 4-5 = yes). I analyzed the results from this scoring model using both exact agreement and Cohen's kappa calculations. In doing so, I maintained the midpoint rating of 3 as uncertain, thus to be disregarded in analyzing reliability of the dichotomous yes/no model of response.

The results of the analysis of rater re-rater agreement by item and for the entire instrument may be seen in Table 3.1.

Table 3.1 Rater Re-Rater Results from Formative Assessment Instrument

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item A1	87.5%	—	—	100%
Item A2	87.5%	—	—	100%
Item A3	87.5%	—	—	100%
Item A4	75.0%	—	—	100%
Item B1	75.0%	—	—	100%
Item B2	37.5%	—	—	87.5%
Item B3	100%	—	—	100%
Item B4	75%	—	—	100%
Item B5	87.5%	—	—	87.5%
Item C1	62.5%	—	—	87.5%
Item C2	62.5%	—	—	100%
Item C3	100%	—	—	100%
Item C4	87.5%	—	—	100%
Item C5	87.5%	—	—	100%
Item D1	75%	—	—	100%
Item D2	87.5%	—	—	87.5%
Item D3	100%	—	—	100%
Item E1	100%	—	—	100%
Item E2	62.5%	—	—	100%
Item E3	100%	—	—	100%
Total	81.9%	.75	.82	97.5%

As seen in Table 3.1, exact rater re-rater agreement for the instrument as a whole was 131/160, or 81.9% (Cohen's kappa = .75, weighted kappa = .82), indicating substantial to almost perfect agreement over time. The percentages of exact agreement by item ranged from a low of 37.5% to a high of 100%, with a total average exact agreement rate of 60.7%. It can be common in evaluating rater agreement, however, to also consider the degree of agreement of raters within one point of each other (Reynolds et al., 2009). When examined from the perspective of agreement within one point, the agreement rates climb to very high levels. Rater re-rater agreement within one point was 156/160, or 97.5%, with individual items ranging from a low of 91.5% to a high of 100%. This supports confidence in the reliability of this instrument in repeated use by the same observer over time.

As may be seen in Table 3.1f, rater re-rater agreement for the instrument as a whole using the dichotomous yes/no model resulted in an exact agreement percentage of 106/108, or 98.1% (Cohen's kappa = .92). This indicates almost perfect agreement in reliability using a yes/no model.

To evaluate whether rater re-rater agreement differed for individual scales, I calculated agreement rates, kappa coefficients, and weighted kappa coefficients for each of the five scales.

Table 3.1a Rater Re-Rater Results: Scale A "Learning Targets"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item A1	87.5%	—	—	100%
Item A2	87.5%	—	—	100%
Item A3	87.5%	—	—	100%
Item A4	75.0%	—	—	100%
Scale A	84.4%	.75	.82	100%

As seen in Table 3.1, the exact rater re-rater agreement for Scale A was 27/32, or 84.4% (Cohen's kappa = .75, weighted kappa = .82), indicating a substantial to almost perfect level of agreement over time. The percentages of exact agreement by item ranged from a low of 75% to a high of 87.5%. Rater re-rater agreement within one point was 32/32, or 100%.

Rater re-rater agreement for Scale A using the dichotomous yes/no model resulted in an exact agreement percentage of 25/25, or 100%. This indicates perfect agreement in reliability using a yes/no model.

Table 3.1b Rater Re-Rater Results: Scale B "Monitoring"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item B1	75.0%	—	—	100%
Item B2	37.5%	—	—	87.5%
Item B3	100%	—	—	100%
Item B4	75%	—	—	100%
Item B5	87.5%	—	—	87.5%
Scale B	75%	.66	.74	95%

Table 3.1b shows that exact rater re-rater agreement for Scale B was 30/40, or 75% (Cohen's kappa = .66, weighted kappa = .74), indicating a substantial level of

agreement over time. The percentages of exact agreement by item ranged from a low of 37.5% to a high of 100%. Rater re-rater agreement within one point was 38/40, or 95%, with individual items ranging from a low of 87.5% to a high of 100%.

Rater re-rater agreement for Scale B using the dichotomous yes/no model resulted in an exact agreement percentage of 23/24, or 95.8% (Cohen's kappa = .92). This indicates almost perfect agreement in reliability using a yes/no model.

Table 3.1c Rater Re-Rater Results: Scale C "Feedback"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item C1	62.5%	—	—	87.5%
Item C2	62.5%	—	—	100%
Item C3	100%	—	—	100%
Item C4	87.5%	—	—	100%
Item C5	87.5%	—	—	100%
Scale C	80%	.68	.72	97.5%

Table 3.1c shows that exact rater re-rater agreement for Scale C was 32/40 or 80% (Cohen's kappa = .68, weighted kappa = .72), indicating a substantial level of agreement over time. The percentages of exact agreement by item ranged from a low of 62.5% to a high of 100%. Rater re-rater agreement within one point was 39/40, or 97.5%, with individual items ranging from a low of 87.5% to a high of 100%.

Rater re-rater agreement for Scale C using the dichotomous yes/no model resulted in an exact agreement percentage of 16/17, or 94.1% (Cohen's kappa = .77). This indicates substantial agreement in reliability using a yes/no model.

Table 3.1d Rater Re-Rater Results: Scale D "Self-Assessment"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item D1	75%	—	—	100%
Item D2	87.5%	—	—	87.5%
Item D3	100%	—	—	100%
Scale D	87.5%	.75	.71	95.8%

Table 3.1d shows that exact rater re-rater agreement for Scale D was 21/24, or 87.5% (Cohen's kappa = .75, weighted kappa = .71), indicating a substantial level of agreement over time. The percentages of exact agreement by item ranged from a low of 37.5% to a high of 100%. Rater re-rater agreement within one point was 23/24, or 95.8%, with individual items ranging from a low of 87.5% to a high of 100%.

Rater re-rater agreement for Scale D using the dichotomous yes/no model resulted in an exact agreement percentage of 21/21, or 100%. This indicates perfect agreement in reliability using a yes/no model.

Table 3.1e Rater Re-Rater Results: Scale E "Peer Assessment"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item E1	100%	—	—	100%
Item E2	62.5%	—	—	100%
Item E3	100%	—	—	100%
Scale E	87.5%	.70	.78	100%

Table 3.1e shows that exact rater re-rater agreement for Scale E was 21/24, or 87.5% (Cohen's kappa = .70, weighted kappa = .78), indicating a substantial level of agreement over time. The percentages of exact agreement by item ranged from a low of 62.5% to a high of 100%. Rater re-rater agreement within one point was 24/24, or 100%.

Rater re-rater agreement for Scale E using the dichotomous yes/no model resulted in an exact agreement percentage of 21/21, or 100%. This indicates perfect agreement in reliability using a yes/no model.

Table 3.1f Rater Re-Rater Agreement Results: Yes/No Response Model

	Exact Agreement	Kappa
Scale A	100%	-
Scale B	95.8%	0.92
Scale C	94.1%	0.77
Scale D	100%	-
Scale E	100%	-
Total	98.1%	0.92

In summary, one type of potential error that may interfere with obtaining an accurate sample for measurement is error across time, the rate re-rate question. How consistent over time will an evaluator be in their use of this instrument in evaluating formative assessment use? In other words, how much will a person's evaluation of formative assessment usage with this instrument vary when observing the same lesson twice?

Regarding the rate re-rate question, in this study when instructional sessions were rated twice over time by the same rater, exact rater re-rater agreement for the instrument as a whole was 131/160, or 81.9% (Cohen's kappa = .75, weighted kappa = .82). Rater re-rater agreement within one point was 156/160, or 97.5%. Using the dichotomous yes/no model resulted in an exact agreement percentage of 106/108, or 98.1% (Cohen's

kappa = .92). These results support confidence in the reliability of this instrument in repeated use by the same observer over time.

Reliability Across Raters

Another source of possible error in the reliability of the instrument is the consistency across two different observers of the same session of classroom instruction. In other words, how closely will two people's ratings match when evaluating the same lesson with the instrument?

To evaluate this, I compared two ratings for each of the 47 instructional sessions. The first ratings were the results of my initial use of the instrument and the second rating was from whichever of the two other raters were assigned to each instructional session. All rater responses were collected and entered into an Excel spreadsheet, and I then determined the level of inter-rater agreement between the pairs of scores. To do so, I followed the same pattern of analysis used in investigating the degree of rater re-rater agreement above.

Table 3.2 contains the results of analyzing the percentage of perfect agreement and agreement within 1 point for each item and for the instrument as a whole, as well as providing Cohen's kappa and Cohen's weighted kappa for the instrument as a whole.

Table 3.2 Inter-Rater Agreement Results

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item A1	70.2%	—	—	95.7%
Item A2	59.6%	—	—	95.7%
Item A3	76.6%	—	—	97.9%
Item A4	63.8%	—	—	91.5%
Item B1	53.2%	—	—	95.7%
Item B2	38.3%	—	—	91.5%
Item B3	51.1%	—	—	100.0%
Item B4	51.1%	—	—	91.5%
Item B5	53.2%	—	—	95.7%
Item C1	48.9%	—	—	97.9%
Item C2	53.2%	—	—	97.9%
Item C3	66.0%	—	—	100.0%
Item C4	63.8%	—	—	100.0%
Item C5	53.2%	—	—	100.0%
Item D1	53.2%	—	—	97.9%
Item D2	83.0%	—	—	93.6%
Item D3	100.0%	—	—	100.0%
Item E1	72.3%	—	—	95.7%
Item E2	66.0%	—	—	93.6%
Item E3	63.8%	—	—	97.9%
Total	60.7%	0.48	0.61	94.5%

As seen in Table 3.2, exact inter-rater agreement for the instrument as a whole was 583/960, or 60.7% (Cohen's kappa = .47, weighted kappa = .61), indicating a moderate to substantial level of agreement across raters. The percentages of exact agreement by item ranged from a low of 38.3% to a high of 100%. As previously noted, it can be common in evaluating inter-rater agreement, however, to also consider the less rigorous stand of inter-rater agreement within one point. When examined from the

perspective of agreement within one point, the percentage of inter-rater agreement was much higher. Agreement within one point was 907/960, or 94.5%, with individual items ranging from a low of 91.5% to a high of 100%. While the kappa coefficient of 0.48 only reflects moderate agreement, the weighted kappa coefficient of 0.61 and the high percentage of ratings within one point support the potential for this instrument to be a reliable source for establishing the presence of formative assessment in classroom instruction, which is its purpose.

As can be seen in Table 3.2f, an analysis of inter-rater agreement for the instrument as a whole using the dichotomous yes/no model resulted in an exact agreement percentage of 573/585, or 97.9% (Cohen's kappa = .87). This indicates almost perfect agreement in reliability using a yes/no model.

To evaluate whether inter-rater agreement differed for individual scales, as with the analysis of rater re-rater agreement, I calculated agreement rates, kappa coefficients, and weighted kappa coefficients for each of the five measurement scales. Those can be seen in the tables below.

Table 3.2a Inter-Rater Agreement Results: Scale A "Learning Targets"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item A1	70.2%	—	—	95.7%
Item A2	59.6%	—	—	95.7%
Item A3	76.6%	—	—	97.9%
Item A4	63.8%	—	—	91.5%
Scale A	67.6%	0.44	0.48	95.2%

Table 3.2a shows that exact inter-rater agreement for Scale A was 127/188, or 67.6% (Cohen's kappa = .44, weighted kappa = .48), indicating a moderate level of

agreement across raters. The percentages of exact agreement by item ranged from a low of 59.6% to a high of 97.9%. Inter-rater agreement within one point was 179/188, or 95.2%, with individual items ranging from a low of 91.5% to a high of 97.9%.

An analysis of inter-rater agreement for Scale A using the dichotomous yes/no model resulted in an exact agreement percentage of 151/155, or 97.4% (Cohen's kappa = .33). This coefficient technically indicates only fair agreement in reliability using a yes/no model, but the lowness of the coefficient is due to statistical properties of calculation based on the extreme number of agreed "no" responses (150 out of 155 eligible paired responses by raters were no-no responses). The exact agreement percentage is more representative of the actual degree of agreement on Scale A using the yes/no model.

Table 3.2b Inter-Rater Agreement Results: Scale B "Monitoring"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item B1	53.2%	—	—	95.7%
Item B2	38.3%	—	—	91.5%
Item B3	51.1%	—	—	100.0%
Item B4	51.1%	—	—	91.5%
Item B5	53.2%	—	—	95.7%
Scale B	49.4%	0.30	0.46	94.9%

Table 3.2b shows that exact inter-rater agreement for Scale B was 116/235, or 49.4% (Cohen's kappa = .30, weighted kappa = .46), indicating a fair to moderate level of agreement across raters. The percentages of exact agreement by item ranged from a low of 38.3% to a high of 53.2%. Inter-rater agreement within one point was substantially higher at 223/235, or 94.9%, with individual items ranging from a low of

91.5% to a high of 100%. The higher weighted kappa reflects this level of agreement within one point.

An analysis of inter-rater agreement for Scale B using the dichotomous yes/no model resulted in an exact agreement percentage of 90/97, or 92.8% (Cohen's kappa = .85). This indicates almost perfect agreement in reliability using a yes/no model.

Table 3.2c Inter-Rater Agreement Results: Scale C "Feedback"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item C1	48.9%	—	—	97.9%
Item C2	53.2%	—	—	97.9%
Item C3	66.0%	—	—	100.0%
Item C4	63.8%	—	—	100.0%
Item C5	53.2%	—	—	100.0%
Scale C	57.0%	0.34	0.47	99.1%

Table 3.2c shows that exact inter-rater agreement for Scale C was 134/235, or 57% (Cohen's kappa = .34, weighted kappa = .47), indicating a fair to moderate level of agreement across raters. The percentages of exact agreement by item ranged from a low of 48.9% to a high of 100%. Inter-rater agreement within one point was 233/235, or 99.1%, with individual items ranging from a low of 97.9% to a high of 100%. The higher weighted kappa reflects this level of agreement within one point.

An analysis of inter-rater agreement for Scale C using the dichotomous yes/no model resulted in an exact agreement percentage of 88/89, or 98.9% (Cohen's kappa = .95). This indicates almost perfect agreement in reliability using a yes/no model.

Table 3.2d Inter-Rater Agreement Results: Scale D "Self-Assessment"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item D1	53.2%	—	—	97.9%
Item D2	83.0%	—	—	93.6%
Item D3	100.0%	—	—	100.0%
Scale D	78.7%	0.52	0.57	97.2%

Table 3.2d shows that exact inter-rater agreement for Scale D was 111/141, or 78.7% (Cohen's kappa = .52, weighted kappa = .57), indicating a moderate level of agreement across raters. The percentages of exact agreement by item ranged from a low of 53.2% to a high of 100%. Inter-rater agreement within one point was 137/141, or 97.2%, with individual items ranging from a low of 93.6% to a high of 100%.

An analysis of inter-rater agreement for Scale D using the dichotomous yes/no model resulted in an exact agreement percentage of 125/125, or 100%. This indicates perfect agreement in reliability using a yes/no model.

Table 3.2e Inter-Rater Agreement Results: Scale E "Peer Assessment"

Item	Exact Agreement	Kappa	Weighted Kappa	+/- 1 Agreement
Item E1	72.3%	—	—	95.7%
Item E2	66.0%	—	—	93.6%
Item E3	63.8%	—	—	97.9%
Scale E	67.4%	0.39	0.44	95.7%

Table 3.2e shows that exact inter-rater agreement for Scale E was 95/141, or 67.4% (Cohen's kappa = .39, weighted kappa = .44), indicating a fair to moderate level of agreement across raters. The percentages of exact agreement by item ranged from a

low of 63.8% to a high of 72.3%. Inter-rater agreement within one point was 135/141, or 95.7%, with individual items ranging from a low of 93.6% to a high of 97.9%.

An analysis of inter-rater agreement for Scale E using the dichotomous yes/no model resulted in an exact agreement percentage of 119/119, or 100%. This indicates perfect agreement in reliability using a yes/no model.

Table 3.2f Inter-Rater Agreement Results: Yes/No Response Model

	Exact Agreement	Kappa
Scale A	97.4%	0.33
Scale B	92.8%	0.85
Scale C	98.9%	0.95
Scale D	100%	-
Scale E	100%	-
Total	97.9%	0.87

In summary, a second type of potential error that may interfere with obtaining an accurate sample for measurement is error across person, the inter-rater reliability question. How consistent are different evaluators are in their rating of formative assessment usage when observing the same session of classroom instruction. In other words, how closely will two people's rating match when evaluating the same lesson with this instrument?

Regarding the inter-rater reliability question, in this study when the same instructional sessions were rated by two different raters, exact inter-rater agreement for the instrument as a whole was 583/960, or 60.7% (Cohen's kappa = .48), indicating a

moderate level of agreement across raters. Agreement within one point was 907/960, or 94.5%. Although the kappa number of 0.48 may only reflect moderate agreement, the high percentage of ratings within one point continues to support the potential for this instrument to be a reliable source for establishing the presence of formative assessment in classroom instruction, as does the higher weighted kappa coefficient of 0.61. The 0.61 kappa rating reflects the overall closeness of the ratings given and demonstrates substantial agreement across raters when considering that closeness. In addition, when using the yes/no scoring model, the reliability of this instrument across raters was found to be almost perfect (Cohen's kappa = .87).

The lower kappa number in Scale B was not surprising. That is due to the fact that several items in this scale called for higher levels of subjective judgment than did items in other scales. I will discuss this further in the following chapter. Despite the low kappa coefficient of 0.34, however, the percentage of agreement within one point remained quite high at approximately 95%. The kappa statistic does not account for the closeness of this relationship, which continues to provide evidence for the reliability of the instrument for its intended purpose of identifying formative assessment in practice. The weighted kappa coefficient of 0.46 gives further support for the nearness of the ratings given, even with the nature of the items within this scale. Additionally, when analyzed using the yes/no scoring model, the kappa coefficient for Scale B was found to be 0.85.

The lower kappa number in Scale C again was not surprising. That is due to issues that emerged during the use of this instrument involving the complexity of delineating feedback, particularly in its relation to instruction. I will discuss this further

in the following chapter. Despite the low kappa level, however, the percentage of agreement within one point once again remained quite high at over 99% and the weighted kappa statistic calculated a higher coefficient of 0.47. When analyzed using the yes/no scoring model, the kappa coefficient was found to be 0.95. These provide support for the reliability of the instrument for its intended purpose of identifying formative assessment in practice.

Internal Consistency

A third area of potential error in an observational instrument is that of internal consistency as measurement error may be introduced by imprecision, inaccuracy, or poor scaling of items within an instrument (Hallgren, 2012). An analysis of internal consistency provides an estimate of the equivalence of sets of items from the same test (Kimberlin & Winterstein, 2008). Regarding this observational instrument, the internal consistency question addresses the equivalence of items within each scale and within the instrument as a whole. In other words, how reliably are the items within the instrument as a whole and within each of the five scales equivalently observing and evaluating different aspects of formative assessment?

The most common method for estimating internal consistency is coefficient alpha or Cronbach's alpha (Cronbach, 1951; Kimberlin & Winterstein, 2008). Cronbach's alpha is a statistic that assesses the reliability of a scale based on its internal consistency (Yang & Green, 2011). It is sensitive to measurement error due to content sampling and is a measure of item heterogeneity that can be applied to tests with items that are scored dichotomously or that have multiple values (Reynolds et al., 2009). To calculate the

internal consistency of this instrument, I utilized SPSS 21 to compute a Cronbach's alpha.

In order to gather the data with which to calculate coefficient alpha, I created an Excel spreadsheet containing all of my initial ratings for each observational session. This totaled 47 observational sessions, including two sets observations of 16 teachers and one set of observations of 15 teachers. As I was the only observer to watch and rate all 47 observational sessions, I utilized the results from my ratings to calculate coefficient alpha. This allowed for optimal rater consistency in analyzing internal consistency. In order to provide clarity and confidence in the estimate of internal consistency, I computed a coefficient alpha for three sets of observations and for the average across all three sets of teacher observation scores. The first set of observations was comprised of scores taken from the first observational session for each of the 16 teachers. The second set of observations was comprised of scores taken from the second observational session for each of the 16 teachers. The third set of observations was comprised of scores taken from the third observational session for the 15 teachers who were observed three times. The fourth set was comprised of the average rating by item for each teacher across the three observational sessions. This resulted in a total of four computations of coefficient alpha for this instrument.

The observational instrument contained a total of 20 items divided into five scales: Learning Targets, Monitoring, Feedback, Self-Assessment, and Peer Assessment. I calculated Cronbach's alpha to determine the level of internal consistency for each scale and for the instrument in its entirety. Results may be seen in Table 3.3.

Table 3.3 Internal Consistency: Cronbach's Alpha by Observation and Average

	1st Observation	2nd Observation	3rd Observation	Average
Scale A	0.54	0.72	0.85	0.71
Scale B	0.72	0.63	0.43	0.68
Scale C	0.88	0.85	0.80	0.91
Scale D	0.42	0.44	-0.17	0.40
Scale E	0.90	0.92	0.90	0.91
Total	0.83	0.83	0.87	0.87

A common recommendation for interpreting acceptable levels of coefficient alpha is found in the following cut-off values: 0.70 for scales in the initial level of development, 0.80 for basic research scales, and 0.90 as the minimal level for scales used for clinical purposes and 0.95 as an ideal level for these scales (Nunnally, 1978). However, what constitutes an acceptable value for Cronbach's alpha may depend on the nature of the scale and the number of items included. It has been suggested while 0.8 may be appropriate for cognitive tests, a cut-off value of 0.7 may be more suitable for ability tests and tests dealing with psychological constructs may be expected to fall below 0.7 (Field, 2009).

As seen in Table 3.3, the total overall estimates of scale reliability for the instrument were acceptable. Cronbach's alpha was calculated four times and the coefficient alpha value exceeded 0.80 for the first set of observations (20 items; $\alpha = .83$), the second set of observations (20 items; $\alpha = .83$), and the third set of observations (20 items; $\alpha = .87$). Coefficient alpha also exceeded 0.80 for the average scores across

observations (20 items; α = .87). These values provide confidence in the scale reliability of the instrument as a whole.

Of interest, especially in light of the discussion in Chapter Five regarding the nature of formative assessment, is the internal consistency of the instrument only using Scales A, B, and C. If Scales D and E (dealing with the components of self-assessment and peer assessment) were removed from the instrument, the resulting coefficient alpha for the instrument as a whole remained at .80 or above for the first set of observations (14 items; α = .80), the second set of observations (14 items; α = .81), the third set of observations (14 items; α = .86), and the fourth set of observation averages (14 items; α = .87). This suggests that the instrument provides a reliable basis for identifying formative assessment in practice when only utilizing Scales A, B, and C.

The scales within the instrument differed in their coefficient alpha values. For example, consider the coefficient values for the average rating across observations, which may give the most accurate overall picture of the instrument. Scale C (Feedback), consisting of five items (α = .91), and Scale E (Peer Assessment), consisting of three items (α = .91) both received the highest values. The values for the next two scales dropped approximately 0.2 points, with Scale A (Learning Targets) consisting of four items (α = .71) and Scale B (Monitoring) consisting of five items (α = .68). The lowest value was for Scale D (Self-Assessment), which consisted of three items (α = .40). It should be noted that there was a statistical problem encountered in SPSS in calculating Scale D for the 3rd observation resulting in the reported score of α = -.17. SPSS reported the presence of a 0 variable item as problematic. Consequently, as the mean scores for the three items comprising Scale D for the 3rd observation (D1 = 2.133, D2 = 1.067, D3 =

1.000) aligned closely to the mean scores for the three items comprising Scale D on average (D1=2.219, D2 = 1.104, D3 = 1.000) and as all other coefficient alpha calculations for Scale D were very consistent (α =.42, α =.44, and α =.40), I disregarded the 3rd Observation Scale D score in favor of the Average Scale D score.

That varying levels of internal consistency values existed within the scales of this instrument was not surprising. In each of the scales, the items were designed to reflect an increasing level of pedagogical sophistication in the use of formative assessment. For example, in Scale B, item B1 asks whether teachers make efforts to monitor learning and item B4 asks whether teachers seek to determine the level of student conceptual knowledge. While related, the latter item demands much more sophisticated use of formative assessment. The levels of internal consistency within scales and within the entire instrument should be viewed in light of that design intention. In fact, the degree to which these items within scales maintained the levels of internal consistency was more than expected.

I would also note that some scales had very consistent values for all four sets of calculations, and some did not. For example, Scale E (Peer Assessment) consisting of four items, received very consistent coefficient alpha values of 0.90, 0.92, 0.90, and 0.91. Some scales were not so consistent. For example, Scale B (Monitoring) consisting of five items, receiving coefficient alpha ratings of 0.72, 0.63, 0.43, and 0.68. In either case, the coefficient alpha values for the set of averages across scores appears to be most representative of the instrument; therefore, I will focus upon those values for the purposes of discussion.

The number of items within each of the five scales should be taken into account when evaluating their coefficient values because a challenge in determining the internal consistency of this instrument by scale is the limited number of items within each scale, ranging from 3 items to 5 items. Although coefficient alpha is sensitive to the internal consistency of a scale, it is heavily influenced by the number of items on it.

> Cronbach's alpha is a function of the average intercorrelations of items and the numbers of items in the scale. It is used for summated scales such as quality-of-life instruments, activities of daily living scales, and the Mini Mental State Examination. All things being equal, the greater number of items in a summated scale, the higher Cronbach's alpha tends to be, with the major gains being in additional items up to approximately 10, when the increase in reliability for each additional item levels off. (Kimberlin & Winterstein, 2008, p. 2277)

Thus, a limited and lower number of items will tend to result in lower Cronbach's alpha values.

Yang and Green (2011), in a critique of some uses of coefficient alpha, illustrated this problem by presenting a hypothetical situation where all items have variances of 1, and correlations between all items are uniformly .3. They pointed out that although this set of items had the same degree of internal consistency (i.e., average inter-item correlation of .3), coefficient alpha was .46 for a two-item scale and .82 for a five-item scale (Yang & Green, 2011). Consequently, I would propose that the consistently higher internal consistency coefficient alpha resulting from examining all 20 items of the instrument speaks more clearly to the actual consistency of the instrument than do the

lower coefficient results based in statistical challenges of scales containing only three items, such as D and E.

In summary, a third type of potential error is error of item scaling, the internal consistency question. How internally consistent are the items and results of the instrument in evaluating formative assessment use? In other words, are items within the instrument as a whole and within each component observing and evaluating the same construct? Regarding the internal consistency question, a common observer's ratings were compared across item scores for all 16 teachers four times, once for each set of three observations and once for the average of those three observations. The resulting Cronbach's alpha calculation exceeded 0.80 for the first set of observations (20 items; $\alpha = .83$), the second set of observations (20 items; $\alpha = .83$), and the third set of observations (20 items; $\alpha = .87$). Coefficient alpha also exceeded 0.80 for the average scores across observations (20 items; $\alpha = .87$). These values provide confidence in the scale reliability of the instrument as a whole. Lower coefficient alpha values for individual scales within the instrument can be understood from the intended design of the instrument and the low number of items comprising each of the scales.

Formative Assessment Use

In seeking to determine whether formative assessment is observable in practice, the study developed an observational instrument for identifying formative assessment use. The instrument incorporated five formative assessment components, rating 20 specific items grouped by component. Raters responded to each item using a Likert-type 1-5 scale (where 1 = no evidence of use and 5 = pervasive or highly effective use)

indicating whether and to what degree each item was observed in practice. The responses from each rater for each observation were entered into an Excel spreadsheet.

The primary purpose of this study revolves around the development of an instrument that can answer the question of whether formative assessment is observable in practice; however, the levels and types of formative assessment use observed during this process can be of interest and benefit. Therefore, I am including the following tables showing the findings about formative assessment use that resulted from the use of this observational instrument.

Table 3.4a displays the average rating across teachers and observers for each of the 20 formative assessment items, calculated by taking the sum total of all ratings given by observers for a given item divided by the total number of times that item was rated. The results include the ratings from all teachers, raters, and instructional sessions.

Table 3.4a Average Formative Assessment Use by Item

Formative Assessment Observational Item	Average Rating
A. Learning Targets: Clarifying Learning Intentions and Sharing Criteria for Success	
1. Does the teacher make certain that students understand the learning intentions for the class session?	1.71
2. Does the teacher make certain that students understand the learning intentions for each activity?	1.84
3. Does the teacher provide examples of high and low quality work?	1.31
4. Does the teacher address potential misunderstandings regarding the criteria for success?	1.66
B. Monitoring: Engineering Effective Classroom Discussions, Questions, and Learning Tasks That Elicit Evidence of Learning	
1. Does the teacher make efforts to monitor student learning on an ongoing basis (i.e., minute-to-minute & day-to-day)?	3.32
2. Does the teacher give students a variety of opportunities and methods (e.g., verbal, written, electronic, & visual) to respond to questions?	3.43
3. Does the teacher use effective questioning strategies (e.g., adequate wait time, open-ended questions) to elicit evidence of learning?	3.15
4. Does teacher monitoring seek to elicit evidence from students of both factual/procedural knowledge and of deeper conceptual knowledge?	2.70
5. Does teacher monitoring seek to elicit evidence of whether students can transfer knowledge within and between disciplines/subjects?	2.01
C. Feedback: Providing Feedback That Moves Learners Forward	
1. Does the teacher provide meaningful feedback (i.e., information with which a learner can confirm, add to, overwrite, tune, or restructure understanding) immediately following formal and/or informal evaluations of student progress?	3.01
2. Does the teacher provide accurate feedback that assists learning?	3.08
3. Does the teacher provide feedback in reference to a criterion-based standard, avoiding feedback based in comparison to other students?	2.81
4. Does feedback describe specific areas of needed improvement and suggest alternative strategies for making that improvement?	2.47
5. Does feedback describe specific student strengths and suggest strategies for continued learning in those areas?	1.91
D. Self-Assessment: Activating Students as the Owners of Their Own Learning	
1. Does the teacher give students opportunities to use self-regulatory competencies, such as the ability to accurately assess their own knowledge?	2.04
2. Does the teacher make efforts to develop self-monitoring competencies in students (i.e., meta-cognitive skills)?	1.12
3. Are students making decisions related to their own improvement on the basis of ongoing assessment data (i.e., ownership of learning)?	1.00
E. Peer Assessment: Activating Students as Instructional Resources for One Another	
1. Does the teacher give students opportunities (e.g., discussions, questions, learning tasks) to engage in peer-monitoring?	1.71
2. Does the teacher utilize the results of peer activities to strengthen ongoing assessment of student learning?	1.43
3. Does the teacher utilize peer activities to help students deepen their understanding of common errors and alternative strategies?	1.38

As the data demonstrate, certain specific aspects of formative assessment were much more likely to be utilized than were others. For example, five items received average ratings above 3.0. Three of those items dealt with the monitoring of student learning, specifically asking about the amount of effort being put into monitoring, the variety of methods/opportunities used in monitoring, and the effectiveness of questioning strategies. The two other items receiving an average rating about 3.0 were in the area of feedback, specifically asking whether the teacher provided meaningful feedback and whether the teacher provided accurate feedback.

As mentioned in considering issues of internal consistency, the instrument was designed to look for increasingly sophisticated uses of formative assessment within each scale. In other words, the earlier items were expected to receive higher rating than the later items within each scale. This design expectation held true in most scales. For example, the first two items under Scale B received average ratings of 3.32 and 3.43, the third received 3.15, the fourth received 2.70, and the fifth received 2.01. This scoring, which is representative of all the scales, reflects the intended design of the instrument.

Table 3.4b displays the average rating of each formative assessment component in the observations conducted over the course of this study. The results include the ratings from all teachers, raters, and instructional sessions.

Table 3.4b Average Formative Assessment Use by Scale

Formative Assessment Scale	Average Rating
A. Learning Targets: Clarifying Learning Intentions and Sharing Criteria for Success	1.63
B. Monitoring: Engineering Effective Classroom Discussions, Questions, and Learning Tasks That Elicit Evidence of Learning	2.92
C. Feedback: Providing Feedback That Moves Learners Forward	2.66
D. Self-Assessment: Activating Students as the Owners of Their Own Learning	1.38
E. Peer Assessment: Activating Students as Instructional Resources for One Another	1.51

As the data demonstrate, certain components of formative assessment were much more likely to be utilized than were others. Most likely to be seen was Scale B (Monitoring) with a scale score of 2.92, followed closely by Scale C (Feedback) with a average scale score of 2.66. Least likely to be found is the Scale D (Self-Assessment) with an average score of 1.38.

Also, it was found that some teachers were more likely to use formative assessment components than were other teachers. Their overall average differed by a range of 0.95 points from low to high. That suggests that a diversity of formative assessment use exists. And, it suggests that formative assessment is observable.

From examining the results of formative assessment use, it is evident that observers were able to distinguish moderate and low levels of formative assessment use in such a way that clear patterns of formative assessment use could be seen. It should be noted, however, that no consistently high levels of formative assessment use were found, whether by item, by scale, or by teacher. That observers did not identify consistently high levels of formative assessment use may be attributed to various potential causes

(e.g., school/district curriculum and instruction approach, teacher training, need for additional professional development). However, the purpose of this study was not to investigate the causes for a lack of formative assessment, but rather to identify whether it was being used in practice.

In summary, the observational results from utilizing the instrument provide suggestive results regarding formative assessment use. It appears that some aspects of formative assessment, such as providing accurate feedback, are more pervasive than others, such as the development of student self-monitoring competencies. It appears that some overall formative assessment components, such as feedback, are more commonly utilized than others. And it appears that use of formative assessment is not consistent from one teacher to another, even within the same school or school district.

Summary

In this observational instrument designed to observe formative assessment in practice, three primary sources of potential error were identified that could potentially interfere with obtaining an accurate sample for measurement. These are error across time, error across person, and error of item scaling. In analyzing the data resulting from the use of this instrument, I addressed each of these potential sources of error using quantitative findings. In doing so, the instrument appears to work well for accomplishing the purpose for which it was designed, which is identifying formative assessment in practice. And the results from using the instrument indicate a positive answer to the question being asked in this study. It demonstrated the ability of observers to consistently (across time and person) identify the presence or absence of formative assessment. The instrument showed itself to possess a clearly acceptable reliability of

scale with a coefficient alpha that exceeded 0.80 for the average scores across observations (20 items; $\alpha = .87$). Therefore, formative assessment does in fact appear to be observable in practice. And the findings indicate that this instrument is a reliable resource for observing and identifying formative assessment in practice.

CHAPTER FIVE: DISCUSSION AND CONCLUSIONS

This chapter includes a brief review of the rationale behind the study and consideration of the significance of the findings. It also includes a discussion of the nature of formative assessment and observations about the core components of formative assessment. In addition, the limitations of the study and implications for future research are considered.

Significance of the Study

The question posed at the beginning of this study was: Is formative assessment observable in practice? This question matters because substantial claims have been made regarding the influence of formative assessment on student learning. From the previously examined works of Paul Black, Dylan Wiliam, and others (Black, Harrison, Lee, Marshall, & Wiliam, 2004; Black & Wiliam, 1998; Wiliam, 2010) to the massive meta-meta analysis of John Hattie (2009), researchers have claimed that formative assessment is one of the most powerful factors in supporting student achievement. The key underlying issue, however, is this: If researchers cannot be confident whether and to what degree formative assessment is present in instruction, then how can they make claims with any confidence regarding the efficacy of formative assessment? If it is uncertain whether and to what degree formative assessment is actually being used in practice, then any claims regarding its influence are difficult to support.

The ultimate contribution of this study is to provide a vehicle through which researchers can make stronger, more substantiated reports about the presence and impact of formative assessment in classroom instruction. A viable method to ascertain whether formative assessment is present is to actually see it in action. Thus, an observational instrument is an appropriate vehicle for identifying the presence of formative assessment. Such an observational instrument can be utilized in viewing classroom instruction to determine whether and to what degree formative assessment is present. As the presence of formative assessment is more clearly identified, researchers of formative assessment may be able to report with more clarity and reliability the impact formative assessment has on the student learning.

There are very practical implications for increasing clarity and confidence in the efficacy of formative assessment for student learning. Because of the resources that have been and will be invested in improving formative assessment, it matters whether formative assessment truly influences student learning as claimed. School administrators and other stakeholders (e.g., parents, politicians, and tax-payers) have a vested interest in the ways in which limited resources are expended in education. Since those resources of time, money, and energy are finite, their expenditure is best used on those things that most influence student learning. Therefore, knowing whether and to what degree formative assessment truly impacts student learning can have a dramatic effect on choices made in the distribution of those resources. These choices include teacher training, professional development, curricular design, and teacher evaluation. And a way to gauge whether and to what degree formative assessment affects student learning is by having a method of identifying it in practice.

Clearly, an instrument for establishing whether formative assessment is present in classroom instruction can be a valuable tool for researchers, teachers, school administrators, and other educational stakeholders. Does such an instrument already exist? An investigation into teacher training and evaluation programs revealed that formative assessment practices can be found embedded in both. However, that investigation showed that there remained an absence of an instrument/method specifically intended to identify its use. As a result of the absence of such an instrument and in light of the potential impact of formative assessment on student learning, the purpose of this endeavor became to develop such an instrument. Thus, in this study, I attempted to determine whether formative assessment could be observed in practice through the creation and implementation of an observational instrument designed for that purpose.

Is formative assessment observable in practice? Although these are initial findings using a limited number of participants, the results of this study suggest that the answer may be yes. Through the use of an observational instrument identifying and evaluating it, observers were able to identify formative assessment as a factor in instruction, teasing it out from instructional practice in general. Observers were able to identify formative assessment and evaluate its use with agreement over time and across raters. The instrument used to guide those observations was shown to have a high degree of scale reliability when viewed in its entirety. Observers were able to distinguish moderate and low levels of formative assessment use in such a way that clear patterns of formative assessment use could be seen. Therefore, when formative assessment is operationalized into specific components and when those components are deconstructed into specific items, it appears that formative assessment can be observed. When those

observational items are then delineated in a protocol of descriptors for rating values and when individuals are trained in understanding those components, items, and descriptors, then it appears that formative assessment use can be observed and evaluated with confidence.

If the test of the presence of formative assessment is whether raters can actually identify it in practice, then formative assessment appears to be observable when present. And if it is observable, then it can be evaluated as a factor in student learning. So long, however, as it remains a vague and undefined construct, then the presence of formative assessment cannot be confidently identified for research or professional development.

Nature of Formative Assessment

One of the key issues in researching and discussing formative assessment has been the lack of clarity regarding its definition. As discussed in Chapter 1, there has been inconsistency regarding such issues as how broadly or narrowly to define formative assessment, whether it is a product or a process, and the time frame within which formative assessment occurs. For the purposes of this dissertation, I relied primarily on the definition published in 2007 by a Council of Chief State School Officers (CCSSO) sub-entity, the Formative Assessment for Students and Teachers – State Collaborative on Assessment and Student Standards (FAST SCASS), which stated: "Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (CCSSO, 2012).

During the process of developing, utilizing, and evaluating the instrument to observe formative assessment in practice, I became more cognizant of the need to further

clarify the definition of formative assessment. The study led me to envision two primary areas in which an understanding of the nature of formative assessment may become more focused. The first area addresses the issues of timing and function of formative assessment within the larger picture of instruction. The second area addresses the issue of operationalizing formative assessment around its key components. The need for a more clear and focused understanding of the nature of formative assessment has less to do with the CCSSO definition than it does with the ways in which that definition may be operationalized as a factor of influence in student learning.

I would define formative assessment as *a dynamic interchange between teacher and student in which instruction is adapted continuously based on student learning status*. This definition speaks to the first area in which the nature of formative assessment may be clarified. That is the area of the temporal placement of formative assessment within the overarching process of instruction. In seeking to observe formative assessment in action, it became clear that formative assessment is a dynamic process that happens during instruction. Regardless of planning or instrumentation, it is in that dynamic interchange of assessing and responding to student learning progress toward understood goals that formative assessment occurs. As such, to incorporate other facets into the conception of the nature of formative assessment will only breed confusion and diffuse clarity regarding the influence and function of formative assessment.

Defining formative assessment in this way distinguishes it from the day-to-day process of planning that is frequently included in descriptions of formative assessment (Leahy et al., 2005; Thompson & Wiliam 2008). Those descriptions incorporate into formative assessment the planning that teachers make before an instructional session

when that planning is influenced by antecedent assessments of student learning. The planning may be influenced by fundamentally summative assessments of learning (e.g., End of Unit Tests or State Standardized Tests) or assessments done with intentionally formative purposes in prior classes. In either case, I would agree that those assessments could be used profitably in preparing for instructional sessions. I would propose, however, that such a practice should not be termed *formative assessment*, for to do so dilutes the meaning of the term and makes a clear understanding of its role in student learning difficult to delineate. Perhaps, we would be better served to frame the entirety of the instructional process around the dynamic of formative assessment, but distinguishing the distinct components of that process. Such a process, perhaps called Formative Assessment Based Instruction, could incorporate components such as a flexible curriculum, assessment-based planning, teaching that emphasized formative assessment, and an emphasis on self and peer assessment components. The key, however, for truly making formative assessment into an observable, distinctive part of student learning is that it be understood only in reference to the dynamic interchange that happens in the moment between teacher and student.

The role of self and peer assessment components play into the second area in which an understanding of the nature of formative assessment can become more focused. In defining formative assessment as *a dynamic interchange between teacher and student in which instruction is adapted continuously based on student learning status*, I am also seeking to focus its nature operationally. In both operational descriptions of formative assessment that I considered in developing this instrument (see Chapter Two), self and peer assessment were considered critical or key features of formative assessment (Leahy

et al., 2005; McManus, 2008; Thompson & Wiliam 2008). In developing and utilizing the instrument to observe formative assessment in practice, it became clear that the essence of formative assessment rests in evaluating student learning progress in light of commonly understood targets and in providing appropriate feedback to help them progress toward those target. The components of self-assessment and peer assessment, though unquestionably important factors in student learning, are not innately part of the formative assessment process. Instead, they may be instrumental in helping the core operational components of formative assessment: understood learning targets, monitoring student learning, and feedback. Perhaps the similar levels of internal consistency of the instrument when only including Scales A, B, and C and when using all five scales are indicative of the extraneous nature of Scales D and E in identifying the presence of formative assessment use.

Limiting the nature of formative assessment to the three core constituent components of understood learning targets, monitoring student learning, and feedback will provide further clarity for identifying and evaluating its use. Instead of including self-assessment as a core component, self-assessment can be understood as a method for accomplishing the three core components. Student ability to self-assess their understanding of learning targets can play a critical role in the teacher's ability to assure that students understand those targets. Student ability to monitor their progress towards those targets can play a critical role in the teacher's efforts to monitor student progress. Seen as a method to support the accomplishment of the core components of formative assessment, self-assessment wields great potential power. The same can be equally true regarding peer assessment. Student interactions regarding their understanding of learning

targets and one another's learning progress can be profoundly beneficial in the teacher's effort to provide clearly understood learning targets and to monitor student learning. Additionally, peer assessment can provide a method for teachers to provide feedback to students by involving students in the process. While self-assessment and peer assessment are not core formative assessment components, they provide powerful methods of accomplishing those core components.

Some might argue to go even further in narrowing the definition and reducing the operational components of formative assessment. For example, might we remove the first operational component of understood learning targets? Are such understood learning targets indispensible for formative assessment to occur? It is a reasonable question because in some pedagogical approaches (e.g., inquiry-based teaching), a teacher will intentionally omit communication of predetermined learning targets. In other words, the teacher knows the intended learning outcomes, but the students intentionally do not. In fact, such a scenario did present itself during the field testing of the observational instrument. In observing that instructional session, however, the teacher continued to monitor learning and provide feedback. Was formative assessment present or not? I would suggest an affirmative answer because of the dynamic interchange that was occurring between teacher and student in which the teacher's instruction was being adapted continuously on the basis of her monitoring of student learning. Nevertheless, I would also suggest that in much classroom instruction, understood learning targets play an important role in establishing a direction and standard for learning success that supports monitoring and feedback. This is especially true if student self-assessment and self-directed learning is an important part of the teacher's pedagogical approach. Perhaps

the inclusion of the first operational component of understood learning targets is a matter of what is generally part of formative assessment versus what is indispensible for formative assessment. In other words, I would posit that while it is possible for formative assessment to occur without the communication of understood learning targets, the importance of those targets for formative assessment in most contexts remains. Thus, while formative assessment may occur at times without that first component, it continues to play an important role in formative assessment use in general.

What of components two and three, student monitoring and feedback? Might one or the other of those components be removed? My response would be negative regarding the removal of either component. The component of monitoring rests at the heart of formative assessment. The assessment of student learning status is the linchpin of the entire process. Without assessment, there is no formative assessment. And it is the third component, the adaptive response to monitoring, that distinguishes the assessment as formative in nature. Without feedback, it is simply assessment. Thus, while formative assessment may occur without the transmission of understood learning targets, it cannot exist without student monitoring and feedback. It is that dynamic interchange in which instruction is adapted that characterizes formative assessment.

Core Formative Assessment Components

Formative assessment as *a dynamic interchange between teacher and student in which instruction is adapted continuously based on student learning status* is comprised of three core operational components: understood learning targets, monitoring student learning, and feedback. These three core components formed the first three components of the observational instrument used in this study. As the core components, these three

require additional discussion regarding observational and interpretational issues that emerged during this study.

The first core component was entitled *Learning Targets: Clarifying Learning Intentions and Sharing Criteria for Success*. That component contained four items within it that asked the following questions:

1. Does the teacher make certain that students understand the learning intentions for the class session?
2. Does the teacher make certain that students understand the learning intentions for each activity?
3. Does the teacher provide examples of high and low quality work?
4. Does the teacher address potential misunderstandings regarding the criteria for success?

Through the process of developing the instrument, training observers, and conducting observations, it became clear that two aspects of this component require particular explanation. The first is that this component entails more than a teacher stating or posting learning objectives or intentions. With this observational instrument, simply stating or posting learning targets would only result in a rating of "2" for item #1 (see Appendix B: Observational Protocol). The essence of this component calls for teachers to clarify and to evaluate student understanding of learning intentions, be they in regard to the overall session (item #1) or a particular activity within the instructional session (item #2). The focus is not on the teacher's presentation of learning targets but on the teacher's efforts to ensure that students understand learning targets.

The second aspect of this component that required clarification was item #4, the potential misunderstandings regarding the criteria for success. The intent for that item was to observe whether the teacher addressed potential misconceptions about what it meant to successfully complete the assignment. The intent for that item was not to observe whether the teacher addressed potential misunderstandings regarding the content of the lesson. The difficulty for raters in making that distinction between potential misconceptions about the assignment and about the content might lead me to revise that item in future use of the instrument.

The second core component was entitled *Monitoring: Engineering Effective Classroom Discussions, Questions, and Learning Tasks That Elicit Evidence of Learning*. That component contained five items within it that asked the following questions:

1. Does the teacher make efforts to monitor student learning on an ongoing basis (i.e., minute-to-minute & day-to-day)?
2. Does the teacher give students a variety of opportunities and methods (e.g., verbal, written, electronic, & visual) to respond to questions?
3. Does the teacher use effective questioning strategies (e.g., adequate wait time, open-ended questions) to elicit evidence of learning?
4. Does teacher monitoring seek to elicit evidence from students of both factual/procedural knowledge and of deeper conceptual knowledge?
5. Does teacher monitoring seek to elicit evidence of whether students can transfer knowledge within and between disciplines/subjects?

In developing the instrument, training observers, and conducting observations, two items in this component posed interpretative challenges. The first was item #3, which asked

whether the teacher used effective questioning strategies. The challenge presented by that item was in agreeing upon a consistent definition of "effective" questioning strategies. Through the training process, we were able to address this challenge; however, in future use of the instrument, greater detail may be added to the observational protocol.

Item #4 in this second core component of Monitoring also posed interpretive challenges. That item asked whether the teacher sought to elicit evidence of both factual/procedural knowledge and of conceptual knowledge. This item posed two challenges. The first was in maintaining clarity that the focus of the items is to be on monitoring rather than on instruction. Item #4 asks if the teacher is assessing whether the student possesses conceptual understanding of the subject matter. It is not asking whether the teacher is providing instruction in conceptual understanding of the subject matter. The first is a monitoring question, the intent of this component as a whole. The second would be an instructional question, which although a valid concern is not what this instrument is intended to observe. In other words, this component generally and this item particularly is aimed at how the teacher is assessing student learning rather than about how the teacher is presenting information.

In considering item #4 of the second core component of Monitoring, I found that it also posed a second challenge. This challenge touched on a question that has been raised by researchers regarding formative assessment. What role does content knowledge play in formative assessment (Bennett, 2011)? In utilizing the instrument, I came to appreciate more deeply the importance of deep content knowledge for effective use of formative assessment. In this particular item, content knowledge plays a significant role.

If a teacher is to monitor whether a student is gaining deeper conceptual knowledge of a subject, then it would seem logical to necessitate that the teacher has a deeper content knowledge of that subject that goes beyond facts and procedures. Indeed, in order for a rater to determine whether the teacher is monitoring for deep conceptual knowledge, it could be important for the person rating the teaching to possess at least a minimal degree of content knowledge as well. For some items in the instrument, the content knowledge of the teacher (and the observer) is not as critical, such as the question of whether a variety of monitoring techniques and opportunities exist (item #2). However, items relating areas such as the monitoring of conceptual knowledge and transfer of knowledge (item #5) illustrate the importance of content knowledge in formative assessment. Additionally, the third core component, which addresses feedback, also can highlight the importance of content knowledge, as such knowledge will be necessary to provide adaptive, meaningful, correct feedback to students.

The third core component was entitled *Feedback: Providing Feedback That Moves Learners Forward*. That component contained five items within it that asked the following questions:

1. Does the teacher provide meaningful feedback (i.e., information with which a learner can confirm, add to, overwrite, tune, or restructure understanding) immediately following formal and/or informal evaluations of student progress?
2. Does the teacher provide accurate feedback that assists learning?
3. Does the teacher provide feedback in reference to a criterion-based standard, avoiding feedback based in comparison to other students?

4. Does feedback describe specific areas of needed improvement and suggest alternative strategies for making that improvement?
5. Does feedback describe specific student strengths and suggest strategies for continued learning in those areas?

In this component, the observational and interpretational issue was not with any particular item. Rather, the primary issue with this component dealt with the nature of feedback itself. In developing the instrument, particularly in the field-testing portion of that process, this question arose repeatedly. How does feedback differ from instruction and how can we delineate the two? Finally, we came to recognize that feedback, which is instructive by nature, always occurs in response to the monitoring of student learning. In other words, feedback is always responsive. It may be planned or unplanned feedback, but regardless, it is a response by the teacher to the recognized learning status of the student. It can be programmed feedback, in which a teacher has predetermined instructional responses to make in response to the progress of a student or students. It can be spontaneous feedback, in which a teacher responds extemporaneously during instructional interactions with students. In either case, however, feedback is always instruction given in response to an evaluation of student progress.

Limitations

Inherent within any research, a common concern is the limitations of the study, which identifies potential weaknesses of the study (Castetter & Heisler, 1984). This study is no exception. Limitations to this study include the limited number of participants (16), the limited number of observations (3), and the limited length of those observations (30 minutes maximum). The involvement of a greater number of teachers and a longer

length of observational time might lend greater strength to the findings. The challenge in involving those greater numbers, however, is resources in terms of the time required, teacher and observer availability, and the necessary funding. The lack of major rating differences across multiple observations of each teacher indicated that the number of observations per teacher was not a serious concern. A related limitation is the number of raters involved in the study. While six raters were trained to use the instrument in observing and evaluating formative assessment use, only three of those raters were able to participate in the observational process due to time and logistical restraints. A greater number of available raters would provide greater strength to the inter-rater reliability findings.

Limitations to this study also include generalizability issues due to the homogeneity of the participants demographically. The teachers involved in the study were all Anglo, teaching in schools within the same school district within one geographical area of the United States. They were almost all female with similar educational backgrounds. While it may not be unexpected for the majority of elementary teachers to be white females, it should be noted. In terms of generalizability limitations, it should also be noted again that due to the structure of the ITML project, the observations were only of class sessions involving mathematics instruction.

Additionally, due to logistical constraints of the study, a combination of in person observations of classroom instruction and later video-taped observations of classroom instruction were used. Occasionally both raters were present in the classroom at the same time due to a teacher's unwillingness to be videotaped, but the vast majority of evaluations of formative assessment use was a combination of in-person and videotaped

observations. I experienced advantages and disadvantages with each method. The advantages of observing in-person was that I could see and hear things not captured on film and could experience firsthand the classroom environment. The disadvantages were the lack of a 'pause button' for taking notes and the challenges of cognitively focusing on multiple observational questions at the same time. The advantages and disadvantages of video-taped observations were exactly the opposite. With videos, I was able to pause, rewind, and re-watch in order to take notes and mentally process the classroom instruction; however, I could not experience the classroom beyond the lens. While the advantages and disadvantages may balance each other, there remains a question of whether a more consistent methodology might provide more reliable results when using multiple raters or if comparing teachers observed with different methods.

Recommendations for Further Research

This study strived to answer the question of whether formative assessment is observable in practice. In seeking to answer that question, I developed an observational instrument for identifying formative assessment in classroom instruction. The use of that instrument resulted in findings that suggest a positive answer to the question of whether formative assessment is observable. Those finding, however, lead to further points of needed inquiry.

The first point of potential research deals with creating and evaluating a revised version of the observational instrument. The revised instrument would be based on the proposed definition of formative assessment as *a dynamic interchange between teacher and student in which instruction is adapted continuously based on student learning status* and on the proposed three core operational components of understood learning targets,

student monitoring, and feedback. Creating a revised instrument on that basis would involve eliminating the two components of self-assessment and peer assessment, possibly incorporating elements of those methodological approaches into the three core components. Additionally, creating such a revised instrument could allow for an increased number of observational items within each scale, which would offer the possibility of an increased level of internal consistency by scale. Thus, the revised instrument could be of more use for quantitative formative assessment research by scale.

The second point of potential inquiry deals with a broader use of the instrument itself. In the future, this observational instrument, or its revised version, could be utilized as a resource to help researchers make stronger, more substantiated reports about the presence and impact of formative assessment. In order to best accomplish that purpose, additional research should be done on the instrument. For example, as this study observed mathematics instruction only, how would the results of the instrument compare when observing instruction in other subjects, e.g., reading? And at the same time, what might that reveal about a teacher's use of formative assessment across subject? Again, as this study observed elementary education only, how would the results of the instrument compare when observing instruction in secondary education or adult education? And what might that reveal about the similarities and differences in using formative assessment with different age groups?

Additionally, as this study observed instruction within a single school district, how would the results of the instrument compare if used in other school districts operating under other curricular models? And what might that reveal about the relationship between curricular choice and formative assessment use? As these examples

illustrate, there are numerous contexts in which the instrument may be tested further. Perhaps even more significantly in regard to understanding formative assessment, utilizing this instrument in those various contexts can provide a wealth of data about formative assessment itself.

A third point of potential inquiry resulting from this study deals with the relationship of self-assessment to formative assessment. It raises a number of questions regarding self-assessment. How does self-assessment support formative assessment? If self-assessment is not to be understood as a component of formative assessment but rather as a method for accomplishing the components of formative assessment, then how does that relationship function? It appears evident that self-assessment can be a powerful force in student learning, and formative assessment claims to be a powerful force in student learning. How, then, do they work together? What components of formative assessment are best accomplished through self-assessment strategies? Would a teaching approach that incorporates, or centers on, self-assessment result in more effective formative assessment use? These and other questions emerge when self-assessment is viewed as a method for accomplishing formative assessment components. Making inquiries into the functional role that self-assessment plays in the formative assessment process could prove most worthwhile in supporting the efficacy of both for student learning.

A fourth point of inquiry resulting from this study deals with the relationship of peer assessment to formative assessment. As with self-assessment, the study raises a number of questions regarding peer assessment. If peer assessment is understood as a method for accomplishing the components of formative assessment, how does that

relationship function for each of the three components of formative assessment: understood learning targets, monitoring student learning, and feedback? In what ways does peer assessment support each of those three components? Does the use of peer assessment for formative assessment purposes reveal those functions and characteristics of peer assessment that can be most beneficial for student learning? As with self-assessment, making inquiries into the functional role of peer assessment could prove most worthwhile.

Conclusion

I came into this study with the following question: Is formative assessment observable in practice? By successfully developing and using an instrument designed to identify formative assessment in classroom settings, I have found evidence suggesting that formative assessment is observable in practice.

I have also discussed that in order for formative assessment to become a quantifiable factor in researching influences on student learning, a narrowing and focusing of its definition is in order. I have suggested that formative assessment be understood as *a dynamic interchange between teacher and student in which instruction is adapted continuously based on student learning status*. The rationale for such a focused definition is not to discover what is observable and then define the construct of formative assessment accordingly. This is not an emasculating of formative assessment into something less so that it can be seen and identified for research. Rather, the more focused definition recognizes and appreciates the crucial and unique character of formative assessment, which is the dynamic interchange between teacher and student that

happens in the classroom. Recognition of that nature may both allow its influence to be studied and its true potential to be reached.

REFERENCES

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034-1037.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*: Brooks/Cole Monterey, CA.

Amazon. (2013). "formative assessment". Retrieved March 28, 2013, from http://www.amazon.com

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261-271. doi: 10.1037/0022-0663.84.3.261

Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80*(3), 260-267. doi: 10.1037/0022-0663.80.3.260

Andrade, H. L. (2010). Summing up and moving forward: Key challenges and future directions for research and development in formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 344-351). New York, NY: Routledge.

Archer, A. L., & Hughes, C. A. (2010). *Explicit instruction: Effective and efficient* teaching. New York, NY: The Guilford Press.

"assessment," (n.d.). *Dictionary.com*. Retrieved from http://dictionary.reference.com/

Baker, S.K., Gersten, R., Haager, D., & Dingle, M. (2006). Teaching practice and the reading growth of first-grade English learners: Validation of an observation instrument. *The Elementary School Journal, 107*(2), 199-219.

Baldwin, M. D., Keating, J. F., & Bachman, K. J. (2006). *Teaching in secondary schools: meeting the challenges of today's adolescent*. Upper Saddle River, N.J.: Pearson/Merrill/Prentice Hall.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*(2-3), 70-91.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. doi: 10.1080/0969594x.2010.513678

Bill & Melinda Gates Foundation, (2010). *Classroom observations and the MET project*

Bill & Melinda Gates Foundation, (2010). *Danielson's framework for teaching for classroom observations*

Bill & Melinda Gates Foundation, (2012). *Gathering feedback for teaching: Research paper*

Bill & Melinda Gates Foundation, (2010). *The CLASS protocol for classroom observations*

Bill & Melinda Gates Foundation, (2010). *The MQI protocol for classroom observations*

Bill & Melinda Gates Foundation, (2010). *The PLATO protocol for classroom observations*

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the Black Box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8.

Black, P., & Wiliam, D. (1998). Inside the Black Box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-144.

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: new roles, new means,* (68th, Part 2 ed., pp. 26-50). Chicago, IL: University of Chicago Press.

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

Board of Regents of the University of Wisconsin System, on behalf of the WIDA Consortium , (2009). *Formative assessment best practices worksheet*. Retrieved from website: http://www.eslportalpa.info/providers/296/FormativeAssessment-Part1-BestPracticesWorksheet.pdf

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245-281. doi: 10.2307/1170684

Castetter, W. B., & Heisler, R. S. (1984). *Developing and defending a dissertation proposal*: Graduate School of Education, University of Pennsylvania.

CCSSO. (2012). Program: Formative assessment for students and teachers (FAST) Retrieved April 21, 2012, from http://www.ccsso.org/Resources/Programs/Formative_Assessment_for_Students_and_Teachers_(FAST).html

Chappuis, S., & Chappuis, J. (2008). The best value in formative assessment. *Educational Leadership, 65*(4), 14-19.

Chappuis, S., & Stiggins, R. J. (2002). Classroom assessment for learning. *Educational Leadership, 60*(1), 40-43.

CLASS™. (2013). Retrieved from http://www.brookespublishing.com/resource-center/screening-and-assessment/class/

The CLASS™ tool. (2013). Retrieved from http://www.teachstone.org/about-the-class/

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213.

Coxon, A. P. M. (1999). *Sorting data : collection and analysis*. Thousand Oaks, Calif.: Sage Publications.

Crisp, G. T. (2012). Integrative assessment: Reframing assessment practice for current and future learning. *Assessment & Evaluation in Higher Education, 37*(1), 33-43. doi: 10.1080/02602938.2010.494234

Croasmun, J. T., & Ostrom, L. (2011). Using Likert-type scales in the social sciences. *Journal of Adult Education, 40*(1), 19-22.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J. (1960). *Essentials of psychological testing*. New York: Harper & Bros.

Danielson, C. (n.d.). *Framework for teaching*. Retrieved from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*: Association for Supervision & Curriculum Development.

Danielson, C. (2011). The Framework for Teaching Evaluation Instrument (2011 ed.). Princeton, N.J.: The Danielson Group.

Danielson, C. (2012). Observing classroom practice. *Educational Leadership, 70*(3), 32-37.

Darling-Hammond, L. (2010). New policies for 21st century demands. In J. Bellanca & R. Brandt (Eds.), *21st Century skills: Rethinking how students learn* (pp. 33-49). Bloomington, IN: Solution Tree Press.

Darr, C. (2005). Hitchhiker's guide to reliability. *SET: Research Information for Teachers, 3*, 59-60.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Los Angeles: SAGE.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation, 14*(7).

Earl, L. M. (2013). *Assessment as learning : Using classroom assessment to maximize student learning*. Corwin

Entwistle, N., & Smith, C. (2002). Personal understanding and target understanding: Mapping influences on the outcomes of learning. *British Journal of Educational Psychology, 72*, 321-321.

Faiks, A., & Hyland, N. (2000). Gaining user insight: a case study illustrating the card sort technique. [Article]. *designing the online help interface for the Cornell University Library gateway system, 61*(4), 349-357.

Field, A. (2009). *Discovering statistics using SPSS*: Sage publications.

Fisher, D., & Frey, N. (2007). *Checking for understanding : Formative assessment techniques for your classroom.* Alexandria, VA: Association for Supervision and Curriculum Development.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American psychologist, 34*(10), 906-911.

Gathering Feedback for Teaching. (2012). Retrieved March 22, 2013, from http://www.coreeducationllc.com/blog2/gathering-feedback-for-teaching/

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education, 57*(4), 2333-2351.

Gliner, J. A., Morgan, G. A., & Harmon, R. J. (2003). Meta-analysis: Formulation and interpretation. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*(11), 1376.

Greenwald, H. J., & O'Connell, S. M. (1970). Comparison of dichotomous and Likert formats. *Psychological Reports, 27*(2), 481-482. doi: 10.2466/pr0.1970.27.2.481

Guilford, J. P., & Guilford, J. (1954). *Psychometric methods* (Vol. 195): McGraw-Hill New York.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology, 8*(1), 23.

Harlen, W. (2007). Criteria for evaluating systems for student assessment. *Studies In Educational Evaluation, 33*(1), 15-28. doi: http://dx.doi.org/10.1016/j.stueduc.2007.01.003

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London; New York: Routledge.

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

Hendry, G. D., Armstrong, S., & Bromberger, N. (2011). Implementing standards-based assessment effectively: incorporating discussion of exemplars into classroom teaching. *Assessment & Evaluation in Higher Education, 37*(2), 149-161. doi: 10.1080/02602938.2010.515014

Heritage, M. (2010). *Formative assessment : Making it happen in the classroom*. Thousand Oaks, Calif.: Corwin.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. . In J. Hiebert (Ed.), *Conceptual and procedural knowledge : The case of mathematics*. Hillsdale, NJ: L. Erlbaum Associates.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., . . . Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment, 17*(2-3), 88-106.

Horn, C., Snyder, B. P., Coverdale, J., Louie, A., & Roberts, L. (2009). Educational research questions and study design. *Academic Psychiatry, 33*(3), 261-267.

Hyeon Woo, L., Kyu Yon, L., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. *Educational Technology Research and Development, 58*(6), 629-648. doi: 10.2307/40929470

Idaho Department of Education, (2012). *Teacher performance evaluation*. Retrieved from website: http://www.sde.idaho.gov/site/teacherEval/

Jacoby, J., & Mattel, M. S. (1971). Three-point Likert scales are good enough. *Journal of Marketing Research (JMR), 8*(4).

Jahrami, H., Marnoch, G., & Gray, A. M. (2009). Use of card sort methodology in the testing of a clinical leadership competencies model. *Health services management research : an official journal of the Association of University Programs in Health Administration / HSMC, AUPHA, 22*(4), 176-183.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*(12).

Jordan, B., & Putz, P. (2004). Assessment as practice: Notes on measures, tests, and targets. *Human Organization : Journal of the Society for Applied Anthropology., 63*(3), 346.

Joyce, B. R., Weil, M., & Calhoun, E. (2011). *Models of teaching*. Boston; London: Pearson Education International.

Kane, M. T. (2006). Validation. *Educational measurement, 4*, 17-64.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23), 2276-2284.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educ. Meas. Issues Pract. Educational Measurement: Issues and Practice, 30*(4), 28-37.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A constructionntegration model. *Psychological Review, 95*, 2, 163-182.

Krathwohl, D. R. (1989). *Social and behavioral science research a new framework for conceptualizing, implementing, and evaluating research studies*. San Francisco, Calif: Jossey-Bass.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership, 63*(3), 18-24.

Leff, S. S., Thomas, D. E., Shapiro, E. S., Paskewich, B., Wilson, K., Necowitz-Hoffman, B., & Jawad, A. F. (2011). Developing and validating a new classroom climate observation assessment tool. *Journal of school violence, 10*(2), 165-184.

Likert, R. (1932). *A technique for the measurement of attitudes*. New York: [s.n.].

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA. *Studies in Second Language Acquisition, 32*(2), 265-302.

Marzano, R. J. (2003). *What works in schools : Translating research into action*. Alexandria, Va.: Association for Supervision and Curriculum Development.

Marzano, R. J. (2006). *Classroom assessment & grading that work*. Alexandria, VA: Association for Supervision and Curriculum Development.

Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? *Journal of Applied Psychology, 56*(6).

Mathematical quality of instruction. (2012). Retrieved from http://isites.harvard.edu/icb/icb.do?keyword=mqi_training

McManus, S. M. (2008). Attributes of effective formative assessment, from http://www.ccsso.org/Documents/2008/Attributes_of_Effective_2008.pdf

McTighe, J., & Seif, E. (2010). An implementation framework to support 21st century skills. In J. Bellanca & R. Brandt (Eds.), *21st Century skills: Rethinking how students learn* (pp. 149-172). Bloomington, IN: Solution Tree Press.

Mevarech, Z., & Fridkin, S. (2006). The effects of IMPROVE on mathematical knowledge, mathematical reasoning and meta-cognition. *Metacognition and Learning, 1*(1), 85-97.

Moreillon, J. (2007). *Collaborative strategies for teaching reading comprehension: Maximizing your impact*. Chicago, IL: American Library Association.

Murayama, K., & Elliot, A. J. (2009). The joint influence of personal achievement goals and classroom goal structures on achievement-relevant outcomes. *Journal of Educational Psychology, 101*(2), 432-447. doi: 10.1037/a0014221

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 149-170.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

Nunnally, J. C. (1978). Psychometric theory: New York: McGraw-Hill.

Perry, N. (2002). Investigating teacher-student interactions that foster self-regulated learning. *Educational Psychologist, 37*(1), 5-15. doi: 10.1207/00461520252828519

Pinchok, N., Brandt, W. C., & Learning Point, A. (2009). *Connecting formative assessment research to practice: An introductory guide for educators*: Learning Point Associates. 1120 East Diehl Road Suite 200, Naperville, IL 60563-1486. Tel: 800-252-0283; Fax: 630-649-6722; Web site: http://www.learningpt.org.

Plato: Protocol for language arts teaching observations. (2009). Retrieved from http://platorubric.stanford.edu/index.html

Popham, J. W. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Ray, J. (1980). How many answer categories should attitude and personality scales use. *South African Journal of Psychology, 10*, 53-54.

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, N.J.: Pearson.

Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education : getting accountability right*. Washington, DC; New York: Economic Policy Institute ; Teachers College Press.

Ruiz-Primo, M. A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies In Educational Evaluation, 37*(1), 15-24. doi: 10.1016/j.stueduc.2011.04.003

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57-84.

Russell, M. K., Airasian, P. W., & Airasian, P. W. (2012). *Classroom assessment : concepts and applications*. Dubuque, Iowa: McGraw-Hill.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education, 28*(2), 147-164.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instr Sci Instructional Science, 18*(2), 119-144.

Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal, 33*(2), 359-382.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and psychological measurement, 64*(2), 243-253.

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

Shapiro, E. S. H. P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools, 41*(5), 551-561.

Shelby, L. B., & Vaske, J. J. (2008). Understanding meta-analysis: A review of the methodological literature. *Leisure Sciences, 30*(2), 96-110. doi: 10.1080/01490400701881366

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5-14.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765.

Stiggins, R. J. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan, 87*(4), 324-328.

Sunderman, G. L., & Kim, J. S. (2007). The expansion of federal power and the politics of implementing the No Child Left Behind Act. *Teachers College Record, 109*(5), 1057-1085.

Thompson, M., & Wiliam , D. (2008). Tight but loose: A conceptual framework for scaling up school reforms. In E. C. Wylie (Ed.), *Tight but loose: Scaling up teacher professional development in diverse contexts* (pp. 1-44). Princeton, NJ: ETS.

Tomlinson, C. A. (2008). Learning to love assessment. *Educational Leadership, 65*(4), 8-13.

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25*(6), 631-645.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20-27.

Tyler, R. W., Gagné, R. M., & Scriven, M. (1967). *Perspectives of curriculum evaluation*. Chicago: Rand McNally.

University of Texas Austin , (n.d.). *The UTeach observation protocol (UTOP) training guide*. Retrieved from website: http://thetrc.org/web/assets/files/evaluation/UTOP_Manual.pdf

Urdan, T. (2004). Using multiple methods to assess students' perceptions of classroom goal structures. *European Psychologist, 9*(4), 222-231. doi: 10.1027/1016-9040.9.4.222

Volante, L., & Beckett, D. (2011). Formative assessment and the contemporary classroom: Synergies and tensions between research and practice. *Canadian Journal of Education, 34*(2), 239-255.

Walsh, J. A., & Sattes, B. D. (2005). *Quality questioning : Research-based practice to engage every learner*. Thousand Oaks, CA; [Charleston, WV]: Corwin Press ; AEL.

WIDA: World-class instructional design and assessment. (2011). Retrieved from http://www.wida.us/index.aspx

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wiliam, D. (2004). *Keeping learning on track: Integrating assessment with instruction*. Paper presented at the The 30th annual conference of the International Association for Educational Assessment (IAEA), Philadelphia, PA.

Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 18-40). New York, NY: Routledge.

Wiliam, D. (2011). What is assessment for learning? *Studies In Educational Evaluation, 37*(1), 3-14. doi: 10.1016/j.stueduc.2011.03.001

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: impact on student achievement. *Assessment in Education: Principles, Policy & Practice, 11*(1), 49-65. doi: 10.1080/0969594042000208994

Wiliam, D., & Thompson, M. (2008). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment : shaping teaching and learning* (pp. 53-82). New York: Lawrence Erlbaum Associates.

Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In J. Dunlosky & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Hillsdale, NJ: Lawrence Erlbaum.

Winne, P. H., & Butler, D. L. (1994). Student cognition in learning from teaching. In T. Husen & T. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 5738-5745). Oxford, England: Pergamon.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich & M. Zeider (Eds.), *Handbook of self-regulation*. San Diego, Calif.: Academic Press.

Wylie, E. C. (2008). *Tight but loose: Scaling up teacher professional development in diverse contexts.* Princeton, NJ: ETS.

Yang, Y., & Green, S. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377-392.

Yeh, S. S. (2006). *Raising student achievement through rapid assessment and test reform*. New York, NY: Teachers College.

Yeh, S. S. (2008). The cost-effectiveness of comprehensive school reform and rapid assessment. *Education Policy Analysis Archives, 16*(13), 1-32.

Zimmerman, B. J., & Schunk, D. H. (2001). *Self-regulated learning and academic achievement: Theoretical perspectives*: Routledge.

APPENDIX A

Formative Assessment Observational Report

Formative Assessment Observation Report

A. Learning Targets: Clarifying Learning Intentions and Sharing Criteria for Success					
	1	2	3	4	5
1. Does the teacher make certain that students understand the learning intentions for the class session?					
2. Does the teacher make certain that students understand the learning intentions for each activity?					
3. Does the teacher provide examples of high and low quality work?					
4. Does the teacher address potential misunderstandings regarding the criteria for success?					

B. Monitoring: Engineering Effective Classroom Discussions, Questions, and Learning Tasks That Elicit Evidence of Learning					
	1	2	3	4	5
1. Does the teacher make efforts to monitor student learning on an ongoing basis (i.e., minute-to-minute & day-to-day)?					
2. Does the teacher give students a variety of opportunities and methods (e.g., verbal, written, electronic, & visual) to respond to questions?					
3. Does the teacher use effective questioning strategies (e.g., adequate wait time, open-ended questions) to elicit evidence of learning?					
4. Does teacher monitoring seek to elicit evidence from students of both factual/procedural knowledge and of deeper conceptual knowledge?					
5. Does teacher monitoring seek to elicit evidence of whether students can transfer knowledge within and between disciplines/subjects?					

C. Feedback: Providing Feedback That Moves Learners Forward					
	1	2	3	4	5
1. Does the teacher provide meaningful feedback (i.e., information with which a learner can confirm, add to, overwrite, tune, or restructure understanding) immediately following formal and/or informal evaluations of student progress?					
2. Does the teacher provide accurate feedback that assists learning?					
3. Does the teacher provide feedback in reference to a criterion-based standard, avoiding feedback based in comparison to other students?					
4. Does feedback describe specific areas of needed improvement and suggest alternative strategies for making that improvement?					
5. Does feedback describe specific student strengths and suggest strategies for continued learning in those areas?					

D. Self-Assessment: Activating Students as the Owners of Their Own Learning					
	1	2	3	4	5
1. Does the teacher give students opportunities to use self-regulatory competencies, such as the ability to accurately assess their own knowledge?					
2. Does the teacher make efforts to develop self-monitoring competencies in students (i.e., meta-cognitive skills)?					
3. Are students making decisions related to their own improvement on the basis of ongoing assessment data (i.e., ownership of learning)?					

E. Peer Assessment: Activating Students as Instructional Resources for One Another					
	1	2	3	4	5
1. Does the teacher give students opportunities (e.g., discussions, questions, learning tasks) to engage in peer-monitoring?					
2. Does the teacher utilize the results of peer activities to strengthen ongoing assessment of student learning?					
3. Does the teacher utilize peer activities to help students deepen their understanding of common errors and alternative strategies?					

APPENDIX B

Observational Protocol

Formative Assessment Observation Protocol

A) Learning Targets: Clarifying Learning Intentions and Sharing Criteria for Success

1) Does the teacher make certain that students understand the **learning intentions** for the **class session**?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not mention learning targets (e.g., objectives, goals) for the class session.
2	N – Superficial or ineffective use	Teacher may post and/or state learning targets , but there is no explanation of what students will need to know
3	N/A - Minimal use or uncertain effectiveness	Teacher describes learning targets (i.e., a specific description of the learning goal being aimed for during the session) adequately in such a way that students can have a clear, solid vision of the learning targets they are responsible for achieving.
4	Y - Frequent Use or Effective	Teacher both describes learning targets and makes clear attempts to evaluate student understanding of those learning targets.
5	Y - Pervasive Use or Highly Effective	Throughout the session, teacher continually reminds class of learning targets, seeks to gauge understanding of those targets, and seeks to evaluate whether students understand them in light of classroom activities.

2) Does the teacher make certain that students understand the **learning intentions** for **each activity**?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Before activities during the class, teacher does not communicate the learning purpose or criteria of learning success for that activity.
2	N – Superficial or ineffective use	Before activities, the teacher may mention a learning purpose for the activity, but there is no explanation other than procedural directions.
3	N/A - Minimal use or uncertain effectiveness	Teacher describes learning targets (i.e., a specific description of the learning goal being aimed for during the activity) adequately in such a way that students can have a clear, solid vision of the learning targets they are responsible for achieving in that activity.
4	Y - Frequent Use or Effective	Before each activity, the teacher both describes learning targets for the activity and makes clear attempts to evaluate student understanding of those learning targets for the activity.
5	Y - Pervasive Use or Highly Effective	Throughout each activity, the teacher continually reminds students of learning targets, seeks to gauge understanding of those targets, and seeks to evaluate whether students understand them in light of the learning intentions for the class session.

3) Does the teacher provide examples of **high and low quality work**?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not provide examples of high or low quality work for the class session or class activities.
2	N – Superficial or ineffective use	Teacher may provide one or two examples of high or low quality work, but with little or no explanation by teacher.
3	N/A - Minimal use or uncertain effectiveness	Teacher may provide one or more examples or high or low quality work, providing explanations regarding the quality. Teacher does not include examples of both low and high quality work. Teacher may provide examples before or during class and/or activities.
4	Y - Frequent Use or Effective	Before the class session and/or individual activities, the teacher provides examples of both high and low quality work, explaining why they are different in quality.
5	Y - Pervasive Use or Highly Effective	Both before and during the class session and/or individual activities, the teacher provides examples of both high and low quality work, comparing and contrasting the reasons that they are different, and evaluating student understanding of those differences.

1) Does the teacher address **potential misunderstandings** regarding the criteria for success?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	The teacher does not address any potential misunderstandings regarding the criteria for success.
2	N – Superficial or ineffective use	Before the class session and/or learning activities, the teacher may mention one or two potential misunderstandings, but the teacher does not explain why they might exist.
3	N/A - Minimal use or uncertain effectiveness	Before the class session and/or learning activities, the teacher describes potential misunderstandings and explains why they might exist on a factual/procedural basis.
4	Y - Frequent Use or Effective	Before the class session and/or learning activities, the teacher describes potential misunderstandings and explains why they might exist both on a factual/procedural basis and on a deeper conceptual basis.
5	Y - Pervasive Use or Highly Effective	Throughout the session and each activity, the teacher makes certain that students understand potential misunderstandings regarding criteria for success and the reasons such misunderstanding might exist.

B) Monitoring: Engineering Effective Classroom Discussions, Questions, and Learning Tasks That Elicit Evidence of Learning

1) Does the teacher **make efforts** to monitor student learning on an ongoing basis (i.e., minute-to-minute & day-to-day)?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher presents material with no evident attempt to evaluate student learning progress.
2	N – Superficial or ineffective use	Teacher presents material with only limited or superficial attempts to evaluate student learning progress.
3	N/A - Minimal use or uncertain effectiveness	Teacher makes periodic attempts to evaluate individual student learning progress during the session, perhaps through asking questions or observing student work, recording student responses, giving and reviewing quizzes, etc.
4	Y - Frequent Use or Effective	Teacher makes frequent efforts to evaluate learning of both individual student progress and of class-wide learning progress (e.g., KWL Chart). Teacher uses various strategies (e.g. rephrasing, clarifying, elaborating, summarizing, and repeating) to confirm assessments.
5	Y - Pervasive Use or Highly Effective	In addition, there is evidence that the teacher is monitoring student learning, either formally or informally, on a day-to-day basis (e.g., pre-testing).

2) Does the teacher give students a **variety** of opportunities and methods (e.g., verbal, written, electronic, & visual) to respond to questions?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not give students opportunities to respond to questions.
2	N – Superficial or ineffective use	Teacher provides students infrequent opportunities to respond and in only one way (e.g., verbal responses).
3	N/A - Minimal use or uncertain effectiveness	The teacher may provide students with frequent opportunities to respond to questions but only one method of response. Alternatively, the teacher may provide infrequent opportunities to respond to questions but does provide alternative methods of response.
4	Y - Frequent Use or Effective	Teacher provides frequent opportunities for students to respond to questions and provides at least two different methods in which they may respond (e.g., individual verbal responses, peer activities, hand signals, manipulatives, electronic response devices, class discussion, etc.).
5	Y - Pervasive Use or Highly Effective	Teacher utilizes questioning throughout class session and provides at least three different methods for students to respond.

3) Does the teacher use **effective** questioning strategies (e.g., adequate wait time, open-ended questions) to elicit evidence of learning?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not use questioning strategies in instruction.
2	N – Superficial or ineffective use	Teacher uses questioning strategies in instruction, but they are ineffective (e.g., yes/no questions, check-listing, answering their own questions) or inaccurate (based in the subject matter).
3	N/A - Minimal use or uncertain effectiveness	Teacher questioning strategies include some effective elements, but questioning also includes ineffective strategies (e.g., leading questions, non-specific questions).
4	Y - Frequent Use or Effective	Teacher uses positive questioning techniques (e.g., adequate wait time, open-ended) but misses opportunities to probe for deeper understanding through questioning. Teacher may only question incorrect responses.
5	Y - Pervasive Use or Highly Effective	Without fail, teacher uses highly effective questioning strategies. Those strategies include follow-up questions, probing questions, etc., that the teacher uses to respond to both incorrect responses and correct responses.

1) Does teacher monitoring **seek to elicit evidence** from students of both factual/procedural knowledge and of deeper **conceptual knowledge**?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher monitoring never moves beyond students' knowledge of facts or procedures.
2	N – Superficial or ineffective use	Teacher monitoring may infer concepts underlying the factual/procedural knowledge, but it does so infrequently and superficially without explanation.
3	N/A - Minimal use or uncertain effectiveness	Teacher will occasionally use a probing question to unearth whether student is gaining deeper knowledge (e.g., "Why do you think so?", "How do you know that?", "What evidence do you have to support your claim?")
4	Y - Frequent Use or Effective	Teacher frequently uses a monitoring approach (e.g., probing questioning, creating practice tests or test items) to elicit evidence of students' conceptual thinking.
5	Y - Pervasive Use or Highly Effective	Throughout the lesson, the teacher consistently seeks to understand whether students are able to move beyond facts and procedures in their thinking, and the teacher uses a variety of strategies for monitoring conceptual learning.

2) Does teacher monitoring **seek to elicit evidence** of whether **students** can **transfer** knowledge within and between disciplines/subjects?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not ask students to demonstrate the ability to connect current learning to other topics within discipline or to other disciplines (e.g., mathematics, science, language arts).
2	N – Superficial or ineffective use	Teacher monitoring may involves real-life applications of knowledge, but it does not evaluate whether students can make explicit connections with other topics or disciplines.
3	N/A - Minimal use or uncertain effectiveness	Teacher monitoring asks students to demonstrate the ability to connect their knowledge with other topics within the discipline (e.g., previous lessons), but it does not ask them to connect their knowledge to other disciplines.
4	Y - Frequent Use or Effective	Teacher monitoring asks students to demonstrate the ability both to connect knowledge to other topics within the discipline and to other disciplines.
5	Y - Pervasive Use or Highly Effective	Teacher monitoring asks students to repeatedly demonstrate the ability both to connect knowledge to other topics within the discipline and to other disciplines, resulting in evidence of a deepening ability in students to transfer knowledge.

C) Feedback: Providing Feedback That Moves Learners Forward

1) Does the teacher provide **meaningful** feedback (i.e., information with which a learner can confirm, add to, overwrite, tune, or restructure understanding) immediately following formal and/or informal evaluations of student progress?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not provide feedback to students, or feedback is primarily in the form of praise, punishment, or extrinsic rewards.
2	N – Superficial or ineffective use	Teacher provides limited feedback that rarely goes beyond "correct/incorrect" or that fails to be clear, purposeful, and meaningful.
3	N/A - Minimal use or uncertain effectiveness	Teacher occasionally responds to monitoring results to provide clear, purposeful, meaningful feedback (e.g., coaching, questioning, correcting, etc.)," to students individually or to the class as a whole.
4	Y - Frequent Use or Effective	Teacher frequently responds to monitoring results to provide clear, purposeful, meaningful feedback to students in a timely nature after formal and/or informal evaluations of student learning progress.
5	Y - Pervasive Use or Highly Effective	Teacher consistently uses a variety of methods (e.g., video, audio, computer-assisted) to provide clear, purposeful, meaningful feedback after formal and/or informal evaluations of student progress.

2) Does the teacher provide **accurate** feedback that assists learning?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not provide feedback to students, or teacher provides feedback that is inaccurate.
2	N – Superficial or ineffective use	Teacher feedback is unclear, misleading, or easily misunderstood by the students.
3	N/A - Minimal use or uncertain effectiveness	Teacher feedback is factually accurate regarding tasks and processes; however, there is no evidence that the feedback is improving student understanding.
4	Y - Frequent Use or Effective	Teacher feedback is factually accurate regarding tasks and processes, and there is evidence that the feedback is improving student understanding, although student growth in understanding conceptual correctness is unknown.
5	Y - Pervasive Use or Highly Effective	There is evidence that accurate feedback is assisting student learning on both a factual/procedural and on a deeper conceptual basis.

3) Does the teacher provide feedback in reference to a **criterion-based standard**, avoiding feedback based in comparison to other students?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not provide feedback to students, or teacher provides feedback that focuses on comparisons with other students.
2	N – Superficial or ineffective use	Teacher feedback primarily focuses on personal student characteristics (e.g., study harder, good job, keep trying) and/or consists largely of written grades without accompanying explanation.
3	N/A - Minimal use or uncertain effectiveness	Teacher responds to monitoring results with feedback that is specifically related to the task or process of learning, but the teacher does not explicitly connect that feedback to class or student learning targets.
4	Y - Frequent Use or Effective	Teacher responds to monitoring results with feedback that is specifically related to the task or process of learning and that explicitly connects the feedback to classroom and/or student learning goals, current progress, and next steps.
5	Y - Pervasive Use or Highly Effective	Teacher uses a variety of methods (e.g., written, verbal, comment only marking) to repeatedly respond to monitoring results with criterion-based feedback that is connected to student learning goals, current progress, and next steps.

1) Does feedback describe specific areas of **needed improvement** and suggest **alternative strategies** for making that improvement?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not provide feedback to students, or teacher feedback is limited to pronouncements of 'incorrect' with no explanation.
2	N – Superficial or ineffective use	Teacher provides correction of incorrect responses (i.e., gives the correct answer), but feedback does not investigate causes for incorrect response. Mention of student improvement needs is primarily limited to personal student characteristics (e.g., you need to try harder, keep working, etc.)
3	N/A - Minimal use or uncertain effectiveness	Teacher responds to monitoring results with feedback that both corrects incorrect responses and investigates the cause of incorrect responses, providing appropriate instruction for subsequent improvement.
4	Y - Frequent Use or Effective	Additionally, teacher feedback provides student with at least two possible strategies for making the needed improvement.
5	Y - Pervasive Use or Highly Effective	Teacher consistently responds to monitoring results with feedback that guides students to identify alternative strategies for strengthening areas of needed improvement, and the teacher does so in connection with identified class and/or student learning targets.

2) Does feedback describe specific **student strengths** and suggest **strategies** for continued learning in those areas?		
Scale	**Attribute**	**Description**
1	N - No evidence of use	Teacher does not provide feedback to students, or teacher feedback is limited to pronouncements of 'correct' with no further amplification.
2	N – Superficial or ineffective use	Teacher provides affirmation of correct responses, but feedback rarely includes suggestions for continued student learning. Mention of student strengths is limited to personal student characteristics (e.g., you're a good worker, you're so smart, etc.)
3	N/A - Minimal use or uncertain effectiveness	Teacher responds to monitoring results with feedback that identifies correct responses, communicates specific student strengths in relation to understood learning targets, and suggests steps for further student learning in that area of strength.
4	Y - Frequent Use or Effective	Additionally, teacher feedback provides student with at least two strategies for continued learning in that area of strength.
5	Y - Pervasive Use or Highly Effective	Teacher consistently responds to monitoring results with feedback that guides students to identify alternative strategies for continued learning in areas of strength, and the teacher does so in connection with identified student and/or class learning targets.

E) Peer Assessment: Activating Students as Instructional Resources for One Another

Scale	Attribute	Description
1)	Does the teacher give students **opportunities** (e.g., discussions, questions, learning tasks) to engage in peer-monitoring?	
1	N - No evidence of use	Teacher does not give students opportunities to assess their own knowledge.
2	N – Superficial or ineffective use	Teacher provides largely superficial opportunities for students to assess their own knowledge (e.g. "Does that make sense?", "Does anyone have a question?") that do not necessitate a response by students.
3	N/A - Minimal use or uncertain effectiveness	Teacher provides some opportunities for students to reflect on their own knowledge (e.g., physical responses, red/green/yellow cards) that require a response from students, but the teacher does not demonstrate a cohesive plan for students to assess their own knowledge.
4	Y - Frequent Use or Effective	Teacher uses peer group activities in which students are explicitly directed to reciprocally identify each others' strengths and areas for improvement in relation to class learning goals.
5	Y - Pervasive Use or Highly Effective	Teacher makes use of multiple group activities in which students reciprocally assess each others' work, share and compare alternative strategies, and track progress over time in relation to individual and class learning targets.

Scale	Attribute	Description
2)	Does the teacher utilize the **results** of peer activities to strengthen ongoing assessment of student learning?	
1	N - No evidence of use	There is no evidence of instruction in skills related to self-monitoring (e.g., self-questioning, strategy selection, etc.)
2	N – Superficial or ineffective use	There is evidence (e.g., charts on the wall) that the teacher is aware of student self-monitoring strategies and skills, but there is no observed instruction in them.
3	N/A - Minimal use or uncertain effectiveness	Teacher mentions and models self-monitoring strategies (e.g., thinks aloud, verbalizes their process of selecting a strategy), but does not engage students in developing those skills. Teacher may ask student to explain why they chose a particular strategy for completing a learning task.
4	Y - Frequent Use or Effective	Teacher provides instruction in self-monitoring skills and strategies and engages students in actively utilizing those skills (e.g. 'Do you agree with this group's answer and why?').
5	Y - Pervasive Use or Highly Effective	Teacher repeatedly instructs, models, and engages students in the use of self-monitoring skills and strategies in relation to established learning targets for the class and/or individual students.

Scale	Attribute	Description
3)	Does the teacher utilize peer activities to help students deepen their understanding of **common errors and alternative strategies**?	
1	N - No evidence of use	Teacher does not use peer activities to deepen understanding of common errors and alternative strategies.
2	N – Superficial or ineffective use	There is evidence that students have personal learning goals, but there is no evidence that current learning/instruction is influenced by those goals.
3	N/A - Minimal use or uncertain effectiveness	There is evidence that students (e.g., using rubrics, exemplars, and work samples) make decisions regarding their own individual progress toward those goals.
4	Y - Frequent Use or Effective	There is evidence that students have established personal learning goals, are aware of their areas of needed improvement, and are engaged in self-directed learning activities related to their improvement needs.
5	Y - Pervasive Use or Highly Effective	Teacher repeatedly uses peer activities that explicitly engage students in identifying common errors and alternative strategies, analyzing the underlying causes of errors and evaluating alternative strategies (e.g., identifying areas of strength to improve a work sample).