THE VALIDATION OF A SHORT-CYCLE FORMATIVE ASSESSMENT OBSERVATION PROTOCOL FOR SCIENCE AND MATHEMATICS INSTRUCTION

Layne Heitz

Dissertation Prepared for the Degree of

DOCTOR OF EDUCATION

UNIVERSITY OF NORTH TEXAS

December 2013

APPROVED:

Colleen Eddy, Major Professor
Pamela Harrell, Committee Member
John Quintanilla, Committee Member
Sarah Pratt, Committee Member
Nancy Nelson, Chair of the Department of Teacher Education and Administration
Jerry R. Thomas, Dean of the College of Education
Mark Wardell, Dean of the Toulouse Graduate School

Over the years, teachers, administrators, and policy makers have been concerned with optimizing learning for all students. The No Child Left Behind Act put an emphasis on summative assessments, which measure what students have learned. In contrast, formative assessment has been shown in many studies to improve student achievement and motivation because it is applied while students are learning. The purpose of this study was to investigate, for middle and high school mathematics and science instruction, the validity and reliability of a newly developed observation instrument called AssessToday, which is used in a single class period to assess a teacher’s use of short-cycle formative assessment.

The content validity of the instrument was supported through an extensive literature review, feedback from experts in the field of formative assessment, and an examination of 98 classroom observations. For assessing reliability of the instrument, inter-rater reliability coefficients were calculated using data collected by trained observers who independently rated teachers during the same class period using three measures: percentage of agreement between raters, Cohen’s kappa, and Fleiss kappa. Cohen’s kappa (N = 36 pairs) ranged from .62 to 1.00 for all observer pairs with an average kappa of .75 for mathematics (n = 16 pairs) and .76 for science (n = 20 pairs). The recommended threshold for kappa is k = .70. An exploratory factor analysis was conducted on the observation data and the determined factors related to the theoretical framework established in the literature. The results affirmed that the instrument is a tool to be utilized in short-cycle formative assessment with middle and high school science and mathematics teachers.
Copyright 2013

by

Layne Heitz
ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for the ability and strength to accomplish this task. I am truly grateful for the many blessings that He has so richly given me.

Second I would like to thank my loving wife April for her patience and support of the long nights and early mornings over the past few years. It would not have been possible to finish this degree without her help and understanding. I would also like to thank my sons, Joshua and Bregan, for their encouragement and confidence in my ability to achieve an advanced degree.

Third, I would like to thank the outstanding faculty that guided me throughout this project and contributed to my understanding of the many constructs involved in educational studies. Dr. Colleen Eddy served as an influential mentor and chair for this project. Her patient instruction and encouragement were greatly appreciated and I could not have succeeded without her. The other members of my committee, Dr. Pamela Harrell, Dr. John Quintanilla, and Dr. Sarah Pratt also significantly contributed to the completion of this study.

Finally, I am thankful to my colleagues and friends who were instrumental in this journey, including Dr. Charles Matthews, chair of the mathematics department at Southeastern Oklahoma State University, B. J. Albright, mathematics teacher at Durant Middle school, Ruth Payne, who spent countless hours in observations and proof reading, Kendall Fleming, editor, and the elders of the 7th and Beech Church of Christ.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>CHAPTER 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Background of U.S. Policy on Education</td>
<td>1</td>
</tr>
<tr>
<td>A Nation at Risk-1983</td>
<td>2</td>
</tr>
<tr>
<td>The Educate America Act-1994</td>
<td>3</td>
</tr>
<tr>
<td>No Child Left Behind-2001</td>
<td>3</td>
</tr>
<tr>
<td>Impact of High-Stakes Testing on Students</td>
<td>4</td>
</tr>
<tr>
<td>Effect of High-Stakes Testing on Teachers</td>
<td>5</td>
</tr>
<tr>
<td>Moving Forward with High-Stakes Testing</td>
<td>5</td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td>6</td>
</tr>
<tr>
<td>Significance of the Study</td>
<td>7</td>
</tr>
<tr>
<td>Definition of Key Terms</td>
<td>8</td>
</tr>
<tr>
<td>CHAPTER 2 LITERATURE REVIEW</td>
<td>11</td>
</tr>
<tr>
<td>Definition of Formative Assessment</td>
<td>11</td>
</tr>
<tr>
<td>Delineation Between Formative and Summative Assessments</td>
<td>11</td>
</tr>
<tr>
<td>The Feedback Loop</td>
<td>12</td>
</tr>
<tr>
<td>Student Involvement in Formative Assessment</td>
<td>12</td>
</tr>
<tr>
<td>Motivational Factors</td>
<td>13</td>
</tr>
<tr>
<td>Constructivism Connection</td>
<td>14</td>
</tr>
<tr>
<td>Constructivism</td>
<td>14</td>
</tr>
<tr>
<td>Exogenous Constructivism</td>
<td>14</td>
</tr>
<tr>
<td>Endogenous Constructivism</td>
<td>15</td>
</tr>
<tr>
<td>Dialectic Constructivism</td>
<td>15</td>
</tr>
<tr>
<td>Vygotskian Constructivism</td>
<td>16</td>
</tr>
<tr>
<td>Formative Assessment and Constructivist Epistemology</td>
<td>17</td>
</tr>
<tr>
<td>Historical Development of Formative Assessment</td>
<td>17</td>
</tr>
<tr>
<td>From 1967 to 1989</td>
<td>18</td>
</tr>
<tr>
<td>Bloom’s Contribution</td>
<td>19</td>
</tr>
</tbody>
</table>
Scores for Ms. Clark ...........................................................................................77
Observation Results with Ms. Reed .................................................................77
Scores for Ms. Reed .........................................................................................78
Data Collection with Ms. Clark ......................................................................78
Observation Results with Mr. Allen ...............................................................79
Scores for Mr. Allen .......................................................................................80
Observation Results with Ms. Reed ...............................................................80
Scores for Ms. Reed .......................................................................................81
Data Collection with Ms. Reed ......................................................................81
Observation Results with Mr. Allen ...............................................................81
Scores for Mr. Allen .......................................................................................82
Observation Results with Ms. Clark ...............................................................82
Scores for Ms. Clark .......................................................................................83
Inter-rater Reliability for Pilot Study ...............................................................83
Summary of Results .......................................................................................84
Conclusions from Pilot Study .........................................................................85
Conclusion .......................................................................................................86
CHAPTER 4 RESULTS .........................................................................................88
Research Question 1 ........................................................................................88
Evidence of Content Validity from Observation Results ...............................95
  Objective Construct for Mathematics and Science ......................................99
  Question Quality Construct for Mathematics and Science .......................99
  Nature of Questioning Construct for Mathematics and Science ................100
  Self-evaluation Construct for Mathematics and Science .........................101
  Student Affect Construct for Mathematics and Science .........................101
  Instructional Adjustment Construct for Mathematics and Science ...........102
  Evidence of learning Construct for Mathematics and Science ...............102
  Evidence for Content Validity .....................................................................103
Research Question 2 .......................................................................................103
Research Question 3 .......................................................................................107
Research Question 4 .......................................................................................116
CHAPTER 1

INTRODUCTION

The improvement of student achievement has been an elusive goal for educators. Over the years, teachers, administrators, and policy makers have been concerned, not only with improving student achievement but also with narrowing the achievement gap by race, ethnicity, language, and special education status (Department of Education, 2008; No Child Left Behind, 2001). Research has demonstrated that one of the most effective strategies teachers can use to accomplish this goal is the intentional use of formative assessment (Black & Wiliam, 1998).

The tool that is discussed in this paper is an observation instrument called AssessToday. This instrument can be used in a single class period to assess a teacher’s use of short-cycle formative assessment. Short-cycle formative assessment is the process of gathering information about a student’s learning and making classroom adjustments based on that information during the learning process. This process has been shown in many studies to improve student achievement and motivation (Black, 2004; Black & Wiliam, 1998; Meisels, Atkins-Burnett, Xue, & Bickel, 2002; Rodriguez, 2004).

Background of U.S. Policy on Assessment in Education

The following section is a brief review of U.S. policies from the past 30 years that have directly influenced the development of its educational system. This section is included to provide a historical framework that will describe the need for this study. In sum, student achievement has been waning. Research-based and proven procedures need to be put in place to ensure the highest level of achievement in U.S. classrooms in order for students to compete in an increasingly global market place.
A Nation at Risk-1983

One of the turning points in American education was the release of the report titled, *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education (NCEE), 1983). Under the direction of T. H. Bell, the Secretary of Education during the Reagan administration, the NCEE was created. This commission was given the task of examining the nation’s educational system and making recommendations for its improvement. Just 15 years after the U.S. landed on the moon, the educational system seemed to be faltering and the public perception about education was declining. The NCEE was given the task of assessing the quality of teaching and learning in our nation’s schools, defining the problems that must be overcome to pursue the course of excellence in our educational system, and making recommendations for improvement (NCEE, 1983).

The findings of the commission were grim. The report published in April of 1983 began with the words, “Our Nation is at risk” (NCEE, 1983, p.3). The report continued to reveal the many failings of the educational system, the decrease in achievement, and the country’s loss of status in the eyes of the rest of the world. The recommendations for improvement from the commission were considerable and they covered almost every aspect of the educational system. The commission recommended that (a) high school graduation requirements should be raised; (b) the standards and expectations of schools and universities should be improved; (c) the amount of time spent at school should be lengthened; (d) teacher requirements for certification should be more stringent; and, (e) financial support for education should be increased. In addition to the various increases, educators and elected officials should be held responsible for achieving the reforms set out in the report, suggesting a new level of political accountability.
The Educate America Act-1994

As a result of the recommendations from the NCEE, many states began to develop standards-based education systems (Department of Education, 2008). These standards-based systems included the writing of content standards, creating textbooks that were aligned with those standards, and writing assessment exams that would measure a student’s mastery of those standards. The federal government became involved by providing funding for the writing and development of the state standards through the Goals 2000: Educate America Act of 1994 (Department of Education, 2008).

In spite of the efforts taken, achievement, as measured by the standardized tests, did not significantly improve as many had hoped. The Third International Mathematics and Science Study (TIMSS) in 1999 brought to light some disturbing news about American students’ achievement in mathematics and science. The evidence presented by Schmidt (1999) suggested that mathematics and science education in the United States were far from satisfactory. The results of the study revealed that eighth grade mathematics achievement in the United States was in 19th place out of the nearly forty nations that were tested. Phillips (2007) reported that only 65% of eighth grade students in the United States scored at or above a basic level of mathematics understanding. The results for science were similar.

No Child Left Behind-2001

With continued low achievement on the standardized tests, the federal government stepped in with a new accountability policy. In January of 2002, a new and more stringent federal policy was signed into law that would change the trajectory of education for the foreseeable future. The reauthorization of the Elementary and Secondary Education Act (1965) retitled No Child Left Behind Act of 2001 (NCLB) put in place a framework of accountability
that would affect every state in the union. NCLB mandated universal literacy and numeracy by 2014 for third through eighth grade levels as measured by standardized testing (Mabry & Margolis, 2006). NCLB required schools to raise the achievement of students each year. The annual yearly progress (AYP) began low but rose quickly and culminated with the requirement that 100% of students in every state must score proficient on state achievement tests (Hill & Barth, 2004).

Unlike the recommendation for testing from the NCEE in 1983, NCLB demanded measured improvement on test scores and attached serious consequences to schools that did not attain them. This propelled high-stakes testing to a new level not experienced before in the education system in the U.S. If schools failed to meet the AYP, they could face various consequences which might include writing improvement plans, replacing teachers and other staff members, or restructuring the school altogether. NCLB brought high-stakes testing and accountability to every student, teacher, and administrator in America’s public schools.

Impact of High-Stakes Testing on Students

Research indicates that high-stakes testing has not had the positive impact some advocates have professed (Amrein, 2003; Byrd-Blake, Afolayan, Hunt, Fabunmi, Pryor, & Leander, 2010; Cizek, 2001; Hill & Barth, 2004; Lee, 2006). High-stakes tests may actually inhibit academic achievement rather than improve it. This is supported by the fact that in some cases the scores on the independent measures decreased after the implementation of state mandated testing (Amrein, 2003). Lee (2006) claims that achievement gains for both mathematics and reading were almost exactly what they were before the high stakes testing of NCLB was enacted. The National Assessment of Educational Progress (NAEP) showed modest gains in mathematics, but reading scores were basically unchanged. The growth pattern in
mathematics after NCLB remains the same as it was before NCLB was implemented (Lee, 2006; Department of Education, 2008). In spite of the fact that the amount of money spent on education per student has almost doubled, the achievement levels, as revealed in the NAEP, have flat-lined (Department of Education, 2008). Student motivation has also declined due to uninteresting classes and fewer elective courses for students to enroll in (Amrein, 2003).

**Effect of High-Stakes Testing on Teachers**

The pressures of high-stakes testing also affect teachers. Black (2001) suggests that the pressure of high student achievement on standardized exams almost inevitably leads teachers to teach only to the test. A substantial number of teachers are teaching to the test and limiting their instruction to only those items they knew would be tested. This includes spending hours memorizing facts and drilling students on test-taking strategies (Amerin, 2003; Byrd-Blake, M., Afolayan, M. O., Hunt, J. W. Fabunmi, M., Pryor, B. W., & Leander, R., 2010; Hill & Barth, 2004; Lipman, 2004). Byrd-Blake et al. (2010) conducted a qualitative study on 42 teachers and found that part of the teachers’ work that they disliked the most was the pressure of test-driven instruction and high-stakes testing. Due to these standardized tests, teachers are less likely to try new teaching strategies, more likely to use a narrow curriculum that they know will be included on the tests, and spend precious class time on test taking strategies instead of the development of deep concepts embedded in the curriculum.

**Moving Forward with High-Stakes Testing**

In spite of the difficulties associated with high-stakes testing and the lack of achievement in our schools, the job of the educator has not changed. High-stakes testing has been part of the educational system for years, and there is no doubt it will be for years to come. Teachers must find a way to fight through the difficulties associated with high-stakes testing and teach students
in the most effective way available. Slavin (1996) and Black (2001) suggested that student achievement will only be changed when teachers adopt more effective instructional methods.

Educators, therefore, are caught in a trap that is spinning them on an endless cycle of teaching to the test so students can pass a single test on a single day in order to satisfy the state for just one more year. The problem with high-stakes testing and achievement is that testing looks at the product instead of the process. The product, achievement, can only be improved by a better process called teaching. As one looks out over the landscape of possibilities for teachers, formative assessment, when used correctly, has shown significant achievement gains for students (Black & Wiliam, 1998; Bloom, 1984; Stiggins & Chappuis, 2005; Wiliam, Lee, Harrison, & Black, 2004). Thus, formative assessment can be part of the process that will lead to increases in student achievement and holds promise for teachers to break free from the cycle of testing, renewing the focus on optimizing student learning.

Purpose of this Study

AssessToday, a newly developed observation instrument, was developed to assist teachers in improving their use of short-cycle formative assessment strategies in the classroom. The purpose of this study is to establish the validity and reliability of AssessToday for science and mathematics. The questions that will guide the study are:

1) Does AssessToday demonstrate content validity as determined by experts in:
   a. the field of formative assessment
   b. professionals in the field of science and mathematics education?

2) Does AssessToday demonstrate reliability in science and mathematics as determined from inter-rater reliability coefficients of trained observers of AssessToday?
3) Does exploratory factor analysis (EFA) identify any underlying factor structures in AssessToday?

4) Does AssessToday display evidence of construct validity as determined by content validity, the EFA, and reliability?

Significance of Study

The current study is significant for students, educators, and researchers. If AssessToday is validated, students will have the opportunity to be taught by a teacher who knows and understands the constructs of formative assessment. Students will be encouraged to become masters of their own learning and become engaged in the process of education. Educators will have the opportunity to improve their teaching practices through proven formative assessment strategies. Researchers will also benefit from this study as the theories of improving a teacher’s use of formative assessment are more clearly defined.

The current study is also significant due to the general design of the constructs. Although the current study only considered inter-rater agreement in science and mathematics, AssessToday has been designed to be applicable in any content area and therefore has the potential to be used in any subject area. Further validity and reliability studies on domain-specific content areas would need to be conducted in order to demonstrate the appropriateness of AssessToday in other subject areas. This can be extremely significant for school systems that are looking for an instrument to be used across curricula.

The creators of AssessToday did not design the instrument as an evaluative tool to be used as judgments on teachers; rather, it is an informative tool to assist teachers in their personal development. Novice teachers, professional learning communities, or experienced teachers who
want to improve their understanding of formative assessment will benefit greatly from the
information and guidance AssessToday can provide.

Definition of Key Terms

Key terms used in the present study are defined below:

• Formative assessment: Formative assessment has been described and defined in various
ways throughout recent literature. Wiliam and Thompson (2007) categorize formative
assessment by the timeframe needed to complete a feedback cycle. Long-cycle formative
assessment takes place over at least one semester, medium-cycle takes place over a few weeks,
and short-cycle takes place in one to two class periods. For the purpose of this study, formative
assessment refers to those activities undertaken by teachers which provide valuable information
that can be used as immediate feedback to modify teaching and learning activities as needed.
These assessments become formative when the evidence collected by the teacher is used to
adjust instruction in order to meet the learning needs of students (Black & Wiliam, 1998).

• Short-cycle formative assessment: Wiliam and Thompson (2007) describe two levels of
short-cycle formative assessment, day-by-day and minute-by-minute. Day-by-day short-cycle
formative assessment is an assessment that takes place over a one or two day period. Minute-by-
minute short-cycle formative assessment takes place during a single class period with a feedback
loop from a few seconds to a few minutes. For the purpose of this study, short-cycle formative
assessment will refer to minute-by-minute formative assessments that usually takes place during
the 45-90 minutes of instruction.

• Feedback loop: The feedback loop is the process in which the teacher gathers evidence
about a student’s learning, interprets the evidence, and uses it to modify what is happening in the
classroom (Black, Harrison, Lee, Marshall, & Wiliam, 2003).
• Summative assessment: Assessments are *summative* when they are used to make evaluative judgments about what has been learned. Examples of when summative assessments take place include the end of a chapter, at the end of a unit of study, or an end of the semester or year (Brookhart, 2010).

• Just-in-time corrections: *Just-in-time corrections* are those corrections made to instruction during instruction (Guskey, 2007). Just-in-time corrections can be pre-planned by the teacher or performed “on the fly” (Heritage, 2007, p. 141). Just-in-time corrections are made as a response to evidence that is collected from students during instruction.

• Instructional adjustments: *Instructional adjustments* are changes a teacher makes to instruction during instruction. Instructional adjustments are made by teachers based on the feedback they receive from students in written or verbal form or by observation of student affect during instruction.

• Learning progression: A *learning progression* is a sequenced set of sub-skills that students must master on their journey to mastering a larger curricular goal (Popham, 2008).

• Grain size: *Grain size* refers to the amount of information a teacher communicates to students in a single setting. Fine grain size would be a small amount of information, whereas course grained would be a large amount of information. Teachers adjust the grain size of their teaching based on the feedback they receive from their students.

• Scaffolding: *Scaffolding* is a teaching strategy where the teacher allows a student to perform a task to the best of their ability. When necessary, the teacher will provide supportive help and additional instruction in order to move the student to the next level of understanding. For this study, scaffolding is based on the zone of proximal development (ZPD), a theory of learning proposed by Lev Vygotsky (Vygotsky, 1962).
• Content validity: The degree to which elements of an assessment instrument are relevant to, and representative of, the targeted construct for a particular assessment purpose (Haynes, 2001)

• Construct validity: Compromises the evidence and rationales indicating the degree to which data from an assessment instrument measures the targeted constructs; includes all evidence bearing on the measure and encompasses all types of validity (Haynes, 2001)

• Inter-rater reliability: The extent of agreement between scores obtained from different assessors (Haynes, 2001)

• Factor analysis: A process in which observed data are used to identify any underlying factor structures.
CHAPTER 2

LITERATURE REVIEW

Definition of Formative Assessment

Many researchers recognize that formative assessment, when used appropriately, can support and enhance learning and motivation (Anthony, 1996; Black & Wiliam, 1998; 2004; Broadfoot & Black, 2004; Shepard, 2000; Stiggins 2006). Because the term formative assessment is used in a very broad way in the literature, a definition is provided for clarity (Black & Wiliam, 2004). The Black and Wiliam (2009) definition of formative assessment is:

Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (p. 6)

This definition has brought to light several characteristics of formative assessment that help define exactly what formative assessment is and the domains categorizing it.

Delineation between Formative and Summative Assessment

First, formative assessment takes place in the classroom and is connected to the next steps in instruction. This helps delineate formative assessment from summative assessment. Summative assessments are designed to promote accountability and separate successful from unsuccessful students (Stiggins, 2006). In most situations summative assessments are not constructed to assist learning but rather designed to evaluate whether, or how much, learning takes place. Formative assessment, in contrast, is designed to assist learning in the classroom. Formative assessment then is not about the giving of grades but a way to generate feedback in order to determining what steps to take next in the teaching process to facilitate the learning process. Wiliam and Thompson (2007) discuss the time frame that is associated with this information loop and suggest that the most effective classroom use of formative assessment is
either short-cycle or medium-cycle assessment loop. The emphasis for this study is on short-cycle formative assessment.

The Feedback Loop

A second item of interest in the definition of formative assessment is that in order for a teacher to know what next steps to take in instruction, evidence of achievement from the students must be elicited, interpreted, and then put into use. Wiliam and Black (1996) describe this process as the feedback loop. Black and Wiliam (1998) suggest that good feedback is at the heart of good pedagogy.

For feedback to be effective, teachers must first gain insights into the students’ prior knowledge of the subject matter (Leahy S., Lyon C., Thompson M., & Wiliam D., 2005). The teacher then must make some judgments about the students’ knowledge and decide what to do next. The teacher’s decision about what to do next hinges on the learning progression and the adjustment triggers the teacher has set in place (Popham, 2008). The learning progression can be broken down into a sequenced set of sub-skills that are used in the design of the learning cycle. Once the teacher understands what the students know, the teacher can begin to assist their students in building additional knowledge. Sadler (1998) describes several of the factors that affect the feedback loop stressing primarily on the students’ involvement in classroom events as well as the expertise of the teacher.

Student Involvement in Formative Assessment

A third characteristic in the definition of formative assessment is that teachers, learners, and their peers should all use the formative assessment process. The previous discussion focused on how/why teachers use formative assessment, but students should also be involved in the formative assessment process. Traditionally, the teacher has been responsible for all aspects of
the classroom environment (Black & Wiliam, 2009). Students, however, need to realize that they are responsible for their own learning. Students need to become self-regulated learners, and this requires them to exert the effort to move their own learning forward. Students need to add self-assessment and peer-assessment to their learning strategies in order to maximize the feedback loops for themselves. Metacognitive practices are also helpful during the learning process. Leahy (2005) suggests that teachers and students should alter the traditional classroom contract by creating an environment of shared responsibility.

In addition to the behavioral characteristics of formative assessment described in the definition of formative assessment, motivational factors also influence students.

Motivational Factors

Student motivation is naturally enhanced through the effective use of formative assessment (Bloom, 1973; Black & Wiliam, 1998; Stiggins, 2010; Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M., Brandon, P. R., Furtak, E. M., Young, D. B., 2008). When formative assessment is used correctly, student engagement and efforts increase because they are able to focus clearly on what teachers want them to learn. More importantly, students learn because they themselves want to learn. Several researchers (McMillan et al., 2010; Miller & Lavin, 2007; Pryor & Torrance, 1997) suggest formative assessment strategies assist students in developing a mastery goal orientation. That is, students pursue learning because they want to learn (intrinsic motivation), not because they want to impress others (extrinsic motivation). Students develop greater confidence in their ability as learners, take more personal responsibility for their learning, develop greater self-regulation and self-evaluation techniques, and show an increase in self-efficacy. They are also motivated through teacher feedback as teachers communicate confidence in their ability to learn. Students also observe the interest and true
Concern emulating from their teachers through the class interaction. These bi-products of formative assessment play a significant role in the learning process.

Constructivism Connection

When developing a theoretical framework for formative assessment, many researchers use constructivism as a descriptive theory for its foundation (Anthony, 1996; Black & Wiliam, 2004; Brookhart, 1997; Pryor & Torrance, 1997). Schunk (2008) argues that constructivism is not a theory but rather an epistemology that provides an explanation for how people learn. Von Glasersfeld (1996) states, “constructivism provides a model of cognition that leads directly to a method of teaching that credits the student with the power to become an active learner” (p. 25). Constructivism, then, is not a theory about teaching but a theory about how students learn.

Constructivism

Proponents of constructivism hold to the position that knowledge is constructed. How that knowledge is constructed, however, is debated among theorists. In a general sense, constructivism consists of three main perspectives: exogenous, the position that the individual constructs knowledge internally through cognition; endogenous, the position that the individual constructs knowledge through interactions socially; and dialectical, the position that the individual constructs knowledge through a combination of cognition and social interaction (Moshman, 1982; O’Donnell, 2012). All constructivist theories fit loosely somewhere along this continuum.

Exogenous constructivism. Exogenous constructivism is characterized by the concept that knowledge is a reconstruction of structures that exist in the environment (Moshman, 1982). Knowledge is constructed through experiences, observation, and modeling. The implication for teaching under this perspective is that the goal of education is to provide clear and
understandable information that the student can internalize. The student is responsible for constructing the new knowledge as it is presented by the teacher. The exogenous perspective emphasizes a direct-teaching model of instruction with the understanding that students must be actively involved in the learning process by internalizing the information presented (O’Donnell, 2012).

Endogenous constructivism. In contrast to the exogenous perspective, where knowledge is constructed from interaction with the outside environment, the *endogenous perspective* suggests that knowledge is created through internal cognitive activity. O’Donnell (2012) suggests that the best example of exogenous constructivism is Piagetian theory. Piaget (1985) describes knowledge creation through a process of assimilation and accommodation. New knowledge is assimilated into already existing schemata which is then modified to accommodate the new information. Students construct new knowledge structures from already existing structures. From this perspective, knowledge creation will follow a predictable sequence. Piaget suggested four stages of cognitive development where a person’s age determines their cognitive ability. The endogenous perspective of teaching emphasizes the development of logical thinking in the classroom through a process of disequilibrium/equilibrium through cognitive conflict. If the cognitive conflict is to create new meaning, the student must assimilate and accommodate the new information in order to resolve the disequilibrium. (O’Donnell, 2012).

Dialectical constructivism. The third perspective of constructivism is *dialectical constructivism*. Dialectical constructivism falls between the two previously mentioned perspectives and holds that knowledge is created from both the environment and interactions between people. O’Donnell (2012) suggests that this type of constructivism is best illustrated by Vygotsky’s theory, which suggests that cognitive processes are first modeled by adults and then
internalized by children. The dialectical perspective of teaching emphasizes interactions between the teacher and the student where the teacher provides models for the students, scaffolds new information, and encourages classroom discourse. Using this perspective, the teacher provides the scientific language and ready-made generalization of concepts to be learned by the student.

Vygotskian Constructivism

Probably the most influential constructionist, whose theories strongly included social influences, is Lev Vygotsky. Vygotsky’s theory is referred to by several different names such as “sociocultural constructivism” (Schunk, 2008, p. 242), “social-cognitive constructivism” (Oxford, 1997, p. 43), “Vygotskian constructivism” (Green & Gredler 2002, p. 57), “social constructivism” (Von Glasersfeld, 1995, p 141), and “Cultural Historical Activity Theory” (Black & Wiliam, 2009, p. 9 ; Pryor & Crossouard, 2008, p. 10). Whichever term is used to describe it, Vygotsky’s theory of learning has had a profound effect on teaching and learning.

Vygotsky has become well-known for his emphasis on social interaction. Green and Gredler (2002) suggest that the exchange between a student and a more experienced learner is the primary mechanism for learning in the Vygotskian approach. Vygotsky recognized that knowledge is gained from interactions with others, both in and outside of the classroom (Vygotsky, 1962). He also believed that knowledge has its origins in the social structure. He was interested in the relationship between language and learning, especially in the social context (Green & Gredler, 2002). In addition to his emphasis on the social aspects of learning, Vygotsky also introduced the concept of the zone of proximal development (ZPD). The ZPD is the area of potential learning. It represents the lower boundary of what can be learned by the student without assistance and the upper boundary of what can be learned through social discourse with a more
experienced learner such as the teacher (Oxford, 1997; Vygotsky, 1962).

Formative Assessment and Constructivist Epistemology

In considering the constructs of formative assessment as described in the definition by Black and Wiliam (2009), formative assessment clearly falls within the realm of constructivism. Specifically, formative assessment places an emphasis on student responsibility through self-assessment, peer-assessment, and active participation in the classroom. This clearly suggests students must construct their own knowledge during the instructional period. It also emphasizes feedback, which involves social interaction with teachers and peers. Formative assessment stresses the importance of the teacher adjusting instruction in accordance with the current level of understanding of the students and using the ZPD with scaffolding to assist students in their learning. Out of the various schisms of constructivism, formative assessment aligns most closely to the Vygotskian model (Pryor & Crossouard, 2008; Black & Wiliam 2009).

Historical Development of Formative Assessment

Student achievement has been on the mind of educators for as long as mankind has been pursuing education. There have always been students who quickly grasped the ideas presented during instruction and appeared to move effortlessly through whatever curriculum is presented. There have also been students, to varying degrees, who struggled with the concepts presented. At one end of the spectrum there have been students who grasped enough knowledge to advance to the next level of learning, while at the other, there have been students who became discouraged and dropped out of formal education altogether. Policy makers and educators have been concerned with this achievement gap and have sought out remedies to close the gap (Bloom, 1974; Guskey, 2005; National Academy of Science, 2007; Stiggins & Chappuis, 2005).
One teaching strategy that has been effective in increasing student achievement, especially among students who have struggled in their academic performances, is formative assessment (Black & Wiliam, 1998; Bloom, 1984; Stiggins & Chappuis, 2005; Wiliam, Lee, Harrison, & Black, 2004). When teachers gather information about the learning of their students, interpret that information, and use that information to make changes in instruction during instruction, they are using formative assessment. These assessments are designed to develop and contribute to students’ intellectual growth.

Black and Wiliam (1998) have presented the most significant evidence about the effectiveness of formative assessment. In their analysis of 250 research studies, the use of formative assessment in classroom innovations produced significant learning gains, with “typical effect sizes of the formative assessment experiments were between .4 and .7” (Wiliam, 2010, p. 20). Wiliam (2010) stressed that these are impressive gains and are larger than those found in most educational studies. Black and Wiliam (1998) also found that the lower-achieving students benefited most from the use of formative assessment. Cizek (2010) states formative assessment “represents the next best hope for promoting greater achievement gains for students” (p. 4). Thus, formative assessment has the potential to close the achievement gap between high and low achievers, which has been an important goal in education (Guskey, 2007; Stiggins, 2009).

From 1967 to 1989

Although the ideas and practice of formative assessment have always been part of the educational process (Wiliam & Black, 1996), the formal development of a classroom structure called formative evaluation started about 45 years ago with Michael Scriven (Scriven, 1967, p. 5). He provided a framework for formative evaluation of curriculum and suggested that it is
more important to evaluate the course while it is being developed than to wait until the finished product was on the market. In evaluating curriculum, he referred to aspects of course revision as “formative” and the final evaluation as “summative” (p. 5). He mentioned that even though his paper discussed the evaluation of curriculum, formative assessments transferred to other kinds of evaluation as well. Bloom (1968) picked up the term “formative” and applied it in a more general sense to the learning process.

Bloom’s contributions. Bloom’s contributions to the development of formative assessment have been important and effective (Guskey, 2005, 2007, 2010). He, along with Hastings and Madaus (1971), described the benefits of using formative assessment practices to assist students in the learning process. For example, they developed concepts such as feedback and correctives that helped students identify what they knew and what they needed to know better.

The formative assessment process was referred to by Bloom as “mastery learning” (Bloom, Hastings, & Madaus, 1971, p. 45). His instructional strategy involved teachers developing learning units and periodically administering a brief formative assessment. Bloom designed tests that provided students with feedback on how well they were progressing through the material. This helped the students realize what they had learned and where they needed to concentrate their efforts as they moved towards mastery of knowledge. Teachers then assigned corrective activities to students as needed for remediation or enrichment for students who had mastered the content of the unit. A second formative assessment was administered after the remediation, giving students a second chance at success.

Bloom found that this method of instruction increased the percentage of students receiving a grade of A on the final exam from 20 percent to 80 percent after the first year and
then to 90 percent after the second year after implementation (Bloom, 1968). He also mentioned that there were significant changes in the ways students viewed themselves and their scholastic abilities. Guskey (2007) states that other improvements were also found in later research, which included “improvements in student’s confidence in learning situations, school attendance rates, involvement in class sessions, attitudes toward learning, and a variety of other affective measures” (p. 24). Because of the great success Bloom experienced through his research, mastery learning blossomed into a nation-wide, and even an international, instructional strategy (Bloom, 1984; Guskey, 2007; Kulik & Kulik, 1989; Kulik, Kulik, & Bangert-Drowns, 1990).

In 1984, Bloom published an article on an extension of his mastery learning strategy. In this study Bloom, along with two doctoral students, compared student learning under three different conditions. The first group was taught with conventional methods (control group) using standard lectures and summative assessments. The second group was taught using Bloom’s mastery learning strategy, and the third group was taught individually by a tutor. The research revealed that the average tutored student scored two standard deviations, or 98% higher than the control group. Similarly, the students who were taught using mastery learning scored one standard deviation above the control group, which is 84% higher than this group. He also reported on six other studies that were conducted, all recording similar results.

The major constructs of Bloom’s work - assessment, feedback, correctives, and enrichment activities - have become major constructs in formative assessment. Perhaps the most important application of his work to formative assessment is the importance of student mindfulness in the learning process. Students know where they are, understand where they are going, and have a clear idea of how to get there.
Scarcity of early research. In spite of the notable success of mastery learning, few research documents examining formative assessment procedures were produced after Bloom’s initial publication (Sadler, 1989). One notable exception was Derek Rowntree (1977) who examined several aspects of classroom assessment including some observations about the use of formative assessment. As to the available literature on assessment, Sadler (1989) stated that, “the vast bulk is concerned with how to use assessment for purposes of grading and ranking. Only a miniscule proportion considers how to use it to enhance the student’s educational growth” (p. 10). One of the main issues he addressed was that assessment can and should be done through avenues other than just testing and giving grades.

Another notable exception that examines formative assessment procedures was Crooks (1988), who examined over 200 research articles dealing with the impact of classroom evaluation practices. He mentioned that too much emphasis was being placed on grades and too little emphasis placed on using evaluation to improve student learning. An emphasis on grading has many undesirable consequences including higher test anxiety, lower motivation and self-efficacy, lower effectiveness of feedback, and poorer social relationships (Amrein, 2003). Assessments should be used in a balanced manner, maximizing the role they can play in student achievement.

Early Construct Development. Although there were few research documents produced involving formative assessment between Scriven’s initial paper in 1967 and the seminal work by Black and Wiliam in 1998, there were several studies conducted about various constructs included in formative assessment. These studies included constructs such as questioning (Crooks, 1988; Redfield & Rousseau, 1981), motivation and student affect (Butler, 1988; Crooks, 1988), and feedback (Brophy & Good, 1984; Johnson & Johnson, 1984; Natrillo, 1987).
Questioning. Questioning was a construct that was prevalent in the literature during the period from 1967 to 1989. One aspect researched was the relationship between a teacher’s use of questioning and student achievement. Several researchers have written summaries of the studies conducted (Crooks, 1988; Levin & Long, 1981; Shiman & Nash, 1974; Wilen & Clegg, 1986). Shiman and Nash (1974) reported that the teachers would ask many questions but ignore questions that would lead to the acquisition of higher-order thinking skills. According to the authors, the use of open-ended questions—questions that probe for deeper understanding and encourage analysis and synthesis of information—should be included in the teaching process in an effort to improve student achievement. Levin and Long (1981) reached similar conclusions in their survey about questioning studies. They found that a number of researchers demonstrated a positive relationship between the level of questions asked by teachers and the actual cognitive process manifested in students’ responses. If higher-level questions were not asked in the classroom, students would not be cognitively challenged to think at a higher level.

The achievement of students is directly related to the questions asked during instruction. Redfield and Rousseau (1981) conducted a meta-analysis of twenty studies that investigated the relationship between teacher questioning and student achievement. The resulting average effect size of .7292 suggested that predominant use of higher-order questions during instruction has a strong positive effect on student achievement. Samson, Strykowski, Weinstein, and Walberg (1987) also conducted a meta-analysis and found an average effect size of .26 favoring the use of higher-order questioning. Although the .26 effect size is not as robust as the .7292 of the previous study, it does indicate a moderate increase in achievement associated with the use of higher-order questions. Good questioning skills, appropriate wait time, and a good mixture of higher and lower level questions are essential to good instruction.
Motivation and student affect. A second construct that was addressed in the literature during this period is that of motivation and student affect. Crooks (1988) states that student interest and motivation to learn are educational outcomes that are at least, if not more important, than cognitive outcomes. His survey of research studies concluded that student motivation and affective measures play a significant role in student achievement. Butler (1988) conducted a study on motivation and student affect. Her study included 132 fifth and sixth grade students and examined the relationship between intrinsic motivation and task-related or ego-related evaluation. Students were assigned one of three feedback conditions: numerical grades only, written comments only, or a combination of grades and comments. Interest and performance were highest among the students who received written comments only and the results were statistically significant (p < .001) when compared to the other two groups. Butler concluded that teacher feedback as part of the evaluation process does indeed affect both interest and motivation.

Brophy and Good (1984) conducted a survey of 33 research studies to examine the relationship between teacher behavior and student achievement. They concluded that students felt secure and confident and were willing to participate and respond to challenge when they were treated with respect and provided with feedback. Students desired an atmosphere that was academically stimulating and demanding. They reported that achievement was higher when the teachers were engaged with the students in the learning process.

Feedback. Another important construct written about during this time period was feedback (Brophy & Good, 1984; Crooks, 1988; Levin & Long, 1981; Natriello, 1987). Ramaprasad (1983) produced an excellent treatise on feedback that has been, and continues to be cited by many formative assessment researchers and scholars. His contribution has primarily
stemmed from his definition of feedback: “Feedback is information about the gap between the actual level and the reference level of a system parameter, which is used to alter the gap in some way” (p. 4). In this definition, there are three important applications to formative assessment.

The first application to formative assessment is that feedback can come from any system parameter. In formative assessment, feedback for the student can come from an outside individual such as the teacher, parents, other students, or from the inside through metacognition and reflection. The second is that there must be a reference level for comparison. For formative assessment to be effective, students must have clear learning targets (Black & Wiliam, 1998; Brookhart, 1997; Stiggins, 2010). Students need to know what is expected of them and what they are required to learn. This leads to the third application of the definition, that feedback must be applied in order to alter the gap (Black & Wiliam, 1998; Popham, 2008; Sadler, 1989). Feedback must be used in the classroom to assist student learning, or it is useless as an instructional tool.

Sadler (1989): A Step Towards Formative Assessment

Sadler (1989) contributed to the change in attitude and focus for research on formative assessment procedures. He noted the great void in formative assessment research and stated:

[D]uring the past 25 years… only cursory attention has usually been given to feedback and formative assessment, and then it is mostly hortatory, recipe-like and a theoretic. In many cases feedback and formative assessment (or their equivalents) are not mentioned at all (p. 122).

He brought many constructs of formative assessment together in order to form an instructional system where assessment was designed to help learning instead of judge learning. He discussed formative assessment constructs including feedback, teacher-student interaction, self and peer assessment, and he stressed the importance of students developing the ability to take control of their own learning and become self-regulated, motivated, independent learners. He mentioned
three necessary conditions for effective classroom assessment: students needing clear targets, the ability of comparing their current level of understanding with a target level, and the ability to take appropriate actions to close the understanding gap. Teachers need to provide appropriate feedback, and help students become self-regulated learners. Also, they need to provide clear and specific targets, consistent scaffolding, and help students assume ownership of their learning environment.

Sadler (1989) made a key statement that is especially important for the development of this study: “A key premise is that for students to be able to improve, they must develop the capacity to monitor the quality of their own work during actual production” (p. 119). The phrase, “during actual production”, is at the heart of this study on formative assessment. The formative assessment that takes place during instruction is crucial for student learning and achievement. This is referred to as “just-in-time” formative assessment (Guskey, 2007, p. 13). Just-in-time formative assessment diagnoses problems in learning, makes needed adjustments in instruction through feedback, and reinforces positive aspects of learning as well. If errors in understanding can be noticed early and changed before they become part of the overall knowledge structure of the student, students will become more effective learners, have greater self-confidence in their ability to learn, and become self-regulated learners who are capable of a lifetime of learning.

From 1990 to 1998

As was noted above, there were several studies about the various constructs of formative assessment but only a few that looked at formative assessment as a framework for teaching. Sadler (1989) mentions only four studies that addressed formative assessment as an appropriate tool for teaching. However, 10 years later, when Black and Wiliam (1998) conducted their
review on formative assessment, they located over 250 studies that applied to formative assessment. Many of the research studies available cite Sadler (1989) as one of their primary sources (Black & Wiliam, 1998; Cowie & Bell, 1999; Pryor & Torrance, 1998), thus building on the arguments established through his work.

Two years prior to the 1998 study, Wiliam and Black (1996) published a paper discussing formative and summative assessments. They provided clarity about formative assessments adding to and enhancing Sadler’s (1989) description that was discussed earlier. Wiliam and Black suggested that formative assessment was nothing new and that “almost all successful teaching relies heavily on adapting the teaching in light of evidence about the success of previous episodes” (p. 538). They suggested that in formative assessments, written and observational evidence were both appropriate for decision-making purposes. They cautioned teachers to realize that evidence can be interpreted correctly and incorrectly, and thus formative assessments should be handled carefully.

Wiliam and Black (1996) also discussed the relationship between formative and summative assessments. The researchers mentioned two main concerns. One was the tension created when the same assessments were used for both purposes. They suggested that formative and summative assessments should be considered two ends of a continuum; formative assessment is used during the teaching and learning process, and summative assessment is used at the end of a teaching cycle. The other concern stemmed from research evidence that showed the difficulty of introducing effective formative assessment into classroom practice.


The study conducted by Black and Wiliam in 1998 changed the way formative assessment was viewed by researchers and practitioners in many places across the globe. Using
the review articles by Natriello (1987) and Crooks (1988) as a baseline, they gathered research studies and reviewed articles from 1988 until 1997 to learn about the formative assessment practices during that time period. They also considered major review articles, journal publications, and reference lists from current articles in order to obtain a thorough review of the literature written from 1988 until 1997. In all, Black and Wiliam located 681 publications that were possible candidates for their study. After reviewing the collected publications, the pertinent studies were narrowed to around 250. Following an extensive reading and review process, the authors coded and labeled the studies, and placed them into seven main categories. Although the authors did not perform a meta-analysis on the combined data, the results from the individual studies revealed that significant learning gains were achieved when teachers incorporated formative assessment practices into their classroom structure.

The contribution of Black and Wiliam (1998) to the development of formative assessment cannot be overstated. Although their contributions to this field of study were many, perhaps most significant were their findings on learning gains: “The research reported here shows conclusively that formative assessment does improve learning. The gains in achievement appear to be quite considerable, and as noted earlier, amongst the largest ever reported for educational interventions” (p. 37). The evidence also revealed that an emphasis on formative assessment in the classroom was particularly beneficial for the lower achieving students.

A second and also important component of this study was the theoretical framework they developed for the use of formative assessment in the classroom. In describing the essential elements of any strategy to improve learning using formative assessment, Black and Wiliam suggested the setting of clear goals, developing appropriate learning tasks, and using feedback to guide instruction. The authors also suggested that the use of self-and peer-assessments should be
considered as part of the essential framework and that a constructivist underpinning should support the entire structure.

The evidence provided by Black and Wiliam (1998) revealed that formative assessment is extremely effective as a teaching/learning tool in the classroom. Their evidence also revealed that significant gains can be achieved through many different routes in the classroom. Formative assessment is a classroom structure that provides a set of guiding principles, i.e. appropriate questioning, attention to student affect, and providing feedback. Specific classroom environment should be decided by each teacher and developed in his or her own way.

The study by Black and Wiliam (1998) suggested that there were also many gaps in the research and development of formative assessment. One significant gap was a basic definition for formative assessment. Many of the studies that they reviewed claimed to be formative assessment but upon review, were something different altogether. Many of the studies had conflicting data or provided inadequate data and were not very useful for drawing any conclusions about formative assessment. They also noticed a lack of detail explaining the difference between summative and formative assessment, as well as clarity about the purpose of the assessments.

From 1998 Forward

As was mentioned, in 1989 when Sadler published his paper on formative assessment, very few research studies or review articles about formative assessment had been conducted. In his article, he mentioned only four studies and stated formative assessment had received minimal attention among researchers. According to Black and Wiliam (1998), the frequency of studies increased between 1989 and 1998 and the focus of assessment had shifted from summative tests,
which were only weakly linked to the learning experiences of students, towards greater interest in the relationship between assessment and learning.

Interest in formative assessment has continued to grow. A search was conducted, (June 2013), using the key words “formative assessment” to ascertain its current frequency of use. The term returned 7,557 studies and articles from a data base of scholarly, peer reviewed publications. A Google search returned 1.8 million web pages which contained the phrase “formative assessment.” The popularity and effectiveness of formative assessment has encouraged many to use the term for actions that are not formative at all. To alleviate this concern, the phrase “assessment for learning” has been adopted by many in the assessment field as a replacement phrase for formative assessment (Wiliam, 2011). A search was also conducted for “assessment for learning” and it returned over 1,600 studies and articles from the aforementioned data base.

Included in the plethora of research articles available today are several research studies that confirm the findings of Black and Wiliam (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Black et al., 2003; Black & Wiliam, 2003; Leahy et al., 2005; Rodriguez, 2004; Wiliam, Lee, Harrison, & Black, 2004) and some that recommend caution when using formative assessment (Prior & Torrance, 1998; Yin et al., 2008). Prior and Torrance (1998), for example, conducted a qualitative study on a single teacher’s use of formative assessment. They analyzed over 100 hours of videotaped classroom interaction and conducted interviews with each student. They concluded that formative assessment use does not guarantee an increase in achievement and motivation. Their conclusions were that:

- There should be a strong connection between formative assessment and the use of learner-centered assessment strategies rather than teacher-centered assessments.
• There should also be an emphasis on the establishment of mastery goals instead of performance goals.

They also suggest the need for consistency between theoretical notions of formative assessment and practical application in the classroom.

Teacher hesitancy with formative assessment. In spite of the evidence to suggest that formative assessment can improve student achievement, teachers have been hesitant to adopt formative assessment practices into their classrooms (Clark, 2011; McMillan, et al., 2010; Stiggins & Chappuis, 2005; Wiliam et al., 2004). There are a number of reasons for this hesitancy. One reason is that the implementation of formative assessment into the classroom takes work and involves a certain amount of change. Most teachers are overworked already and the idea of adding additional duties to an already full schedule does not encourage them to try something new. Black and Wiliam (1998) suggest that teachers need additional education and training to use formative assessment so they can find their own methods of incorporating it into their classroom practices. The pressure of teaching a full year’s curriculum to prepare students for high-stakes testing also discourages the use of formative assessment (Clark, 2011; Wiliam et al., 2004). Clark (2011) suggests that the No Child Left Behind Act (2001) has stifled the development of educational change in the U.S. and has hindered the development of formative assessment as a learning tool in the classroom.

A recent study conducted by McMillan et al. (2010) investigated the question, “[t]o what extent do secondary teachers use different aspects of formative assessment?” (p. 3). The researchers found that most teachers make only limited use of formative assessment practices, in spite of the fact that 60% of the teachers surveyed reported extensive to moderate use of formative assessment in their classrooms. They also suggested that there is a disconnect between
what teachers believe constitutes formative assessment and what is actually supported in the literature. The researchers noted that the primary components that were missing from formative assessment were those that would guide instruction, diagnose student misunderstandings, and provide feedback to the students that would support student learning. These constructs, as discussed earlier, are at the core of formative assessment. McMillan et al. concluded that much work needs to be conducted to enhance teachers’ use of formative assessment.

Refining the field. One of the most important elements in helping teachers feel more comfortable with formative assessment is to narrow the field and provide them with a specific, usable notion of how they can use formative assessment in their classroom. As one looks at a sampling of the material that is now available on formative assessment, it becomes evident that formative assessment has evolved into a rather broad field. Researchers and educators have attached the term *formative assessment* to everything from computer software to professional development programs. Clearly this term has lost some of its meaning since Black and Wiliam (1998) conducted their seminal work. In order to focus formative assessment into a classroom appropriate tool for teachers to use during instruction, the remainder of this literature review will focus on a specific type of formative assessment referred to as “short-cycle formative assessment” (Wiliam & Thompson, 2007, p. 71).

**Time-Scale for Formative Assessment**

One of the ways formative assessment procedures can be a focus for researchers is to consider the amount of time that elapses between an assessment and the accompanying action based on that assessment (Wiliam, 2006). Wiliam and Thompson (2007) suggest three different time frames as seen in Table 1.
The first time frame is referred to as long-cycle formative assessment. Long-cycle formative assessment is used for adjusting instruction across semesters or years. It may involve the evaluation of a course of study or a particular curriculum. Many schools implement professional development based on school-wide assessment information, which is one example of a long-cycle formative assessment. Wiliam and Thompson (2007) suggest long-cycle formative assessments usually last from four weeks to one year.

The second time frame for formative assessment is a medium-cycle formative assessment, which is an assessment used for adjusting instruction during a unit of study. Teachers, for example, may use benchmark tests periodically throughout the chapter and discover students do not understand the concepts presented. If the teacher adjusts the instruction based on the benchmark information, the teacher is using a medium-cycle formative assessment. The time-scale for this type of formative assessment ranges from a few days to four weeks.

A third time frame for formative assessment is short-cycle formative assessment, which is assessment that takes place during the lesson itself. Short-cycle assessments can take place over a one- or two-day time period and is referred to as day-by-day short-cycle assessment.

Table 1

Time Frames For Formative Assessment

<table>
<thead>
<tr>
<th>Type</th>
<th>Focus</th>
<th>Time Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-Cycle</td>
<td>Across learning units, semesters, or years</td>
<td>4 weeks to 1 year or longer</td>
</tr>
<tr>
<td>Medium-Cycle</td>
<td>Within learning units</td>
<td>Several days to 4 weeks</td>
</tr>
<tr>
<td>Short-Cycle</td>
<td>Within and between lessons</td>
<td>A few seconds to 2 days.</td>
</tr>
<tr>
<td>Day-by-day</td>
<td>Over one to two classes</td>
<td>One to two days</td>
</tr>
<tr>
<td>Minute-by-minute</td>
<td>During instruction</td>
<td>A few seconds to one class period</td>
</tr>
</tbody>
</table>

Exit tickets are an example of day-by-day short-cycle assessments since teachers will have to wait until the next class period to correct any misunderstandings. The shortest feedback loop for short-cycle formative assessment takes place in as little as a few seconds or take up to a few minutes and is referred to as *minute-by-minute short-cycle assessment* (p. 71). The key to minute-by-minute short-cycle formative assessment is that the teacher gathers information during instruction and makes adjustments during that class period.

**Short-Cycle Formative Assessment**

According to Wiliam (2006), short-cycle formative assessment has had the most positive effect on student achievement. Leahy, Lyon, Thompson, and Wiliam (2005) also mention that in the longer cycle assessment period the information needed for improving student learning arrives too late. They argue that, “[i]f educators are serious about using assessment[s] to improve instruction, then we need more fine-grained assessment[s], and we need to use the information they yield to modify instruction as we teach” (p. 19). Others in the field agree that the most effective formative assessment takes place during instruction (Brookhart, Moss, & Long, 2008; Cauley & McMillan, 2010; Heritage, Kim, Vendlinski, & Herman, 2009; Popham, 2008; Stiggins, 2009).

*Grain-size* is a phrase often used in the formative assessment literature to refer to the size of the scaffolding or the length of time that is associated with an assessment (Popham, 2008). For example, short-cycle formative assessment is a fine-grained assessment in that the amount of time between the evidence collected and the feedback given is given within a few minutes, whereas long or medium-cycle formative assessment would be considered course-grained.

Names for short-cycle formative assessment. Short-cycle formative assessment is
referred to by different phrases in the literature. Cowie and Bell (1999) refer to this type of assessment as *interactive formative assessment*, which is an assessment that takes place during student-teacher interactions and is impossible to plan for or predict exactly where or when it might happen. Teachers need to notice how students are behaving, their verbal comments, the tone of their discussions, and their body language. They then need to recognize the need for a change and understand the possible instructional strategies at their disposal. The authors stress the importance of teachers accessing the students’ prior knowledge, having a strong pedagogical content knowledge, and possessing mastery of the content being discussed.

Shepard (2000) refers to this type of assessment as *dynamic assessment* (p. 101). She argues that dynamic assessment involves teachers finding out what a student can do on their own, paired with scaffolding content through the ZPD in order to move students forward in their learning. Heritage (2007) refers to short-cycle assessment as “on the fly” assessment (p. 141). She uses the example of a teacher listening to a group discussion and deciding to teach a “pop-up” lesson due to some misunderstandings she heard from the students (p. 141). A “pop-up” lesson is an adjustment in instruction due to an assessment during the class period. Popham (2008) suggests that adjustments in instruction need to be made “on the spot” (p. 11) and follows with the explanation that adjustments need to be made while there is still instructional and learning time available. Finally, Stiggins and Chappuis (2005) state that assessment for learning takes place while students are learning, and its purpose is to provide teachers and students with information needed to facilitate understanding of concepts.

Guskey (2007), in his discourse on Bloom’s mastery learning, uses the phrase “just-in-time correctives” (p. 27). This phrase describes a very important construct in formative assessment; that is, correctives are given at just the right time for maximum improvement. The
most effective use of formative assessment for students is what takes place in the classroom during the instructional period. Timing is very important. At times during instruction students need immediate and specific corrective feedback, while at other times, students should be encouraged to struggle with concepts and the corrective feedback from teachers should be delayed (Brookhart, 2008). This just-in-time correction prevents minor misunderstandings in the content from becoming major barriers later (Guskey, 2010).

Essential elements of short-cycle formative assessment. Many experts in the field of formative assessment have described what are the absolute, critical, and essential elements which must be present in the classroom in order for formative assessment to be effective (Black & Wiliam, 1998; Black & Wiliam, 2003; Black et al., 2004; Cauley & McMillan, 2010; Heritage, 2007; Leahy et al., 2005; McManus, 2008; Stiggins & Chappuis, 2005). Taken together any formative assessment strategy would need to include: (a) clear criteria for success, (b) appropriate learning tasks and classroom discourse, (c) feedback that moves the learning forward, (d) the use of self and peer assessments, and (e) students becoming owners of their own learning. Wiliam (2008) mentions that teachers must tap into these strategies for the full potential of formative assessment to be realized.

The Need for an Observation Instrument

There is ample evidence from research to support the claim that formative assessment, when used correctly and consistently, has the potential to raise student achievement (Black et al., 2004; Black & Wiliam, 1998; Bloom, 1984; Leahy et al., 2005). Many researchers suggest a gap in the current development of formative assessment, specifically, the need to support teachers in their use and development of formative assessment practices. For example, Cizek (2010) stresses the need for field-based techniques that educators can use to achieve the potential
benefits of formative assessment. Further, Wiliam (2010) argues that the most important studies in formative assessment are those studies of how to support teachers in maximizing their use of formative assessment in their own classrooms. Gallagher (2010) clearly states the need, “to develop sensitive instruments to evaluate the assessment quality of formative assessments” (p. 342).

The purpose of this study was to test the validity and reliability of an observation instrument, AssessToday, which was designed to inform teachers on their use of short-cycle formative assessment and provide a platform for improvement. The observation instrument was created by Dr. Colleen Eddy and Dr. Pamela Harrell of the University of North Texas. Although the observation instrument was designed for general use in formative assessment, the current study only examined its use in middle and high school science and mathematics classrooms. Below is a description of the constructs for the observation instrument based on the literature.

The Development of AssessToday

Andrade (2010) suggests formative assessment is used for two overarching purposes: (1) to collect feedback about a student’s learning, and (2) to use this feedback to guide the development of instruction for the process of learning. Teachers and students can use the feedback to guide and develop appropriate decisions about instructional adjustments that take place during instruction. AssessToday, therefore, was developed with these two guiding principles in mind: feedback and instructional correctives. Feedback includes questioning, self-evaluation, and observation of student affect, whereas instructional correctives include instructional adjustments and evidence of learning.
Feedback

One of the most important constructs of AssessToday is feedback (Brookhart, 1997; Cauley & McMillan, 2010; Sadler, 1989). Ramaprasad (1983) defines it as “[i]nformation about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way” (p. 4). Feedback additionally consists of two components: (1) gathering information about the gap and (2) doing something to alter that gap. This is an underpinning of formative assessment. Teachers gather information from their students about their current learning and adjust their teaching accordingly. If the evidence collected from students suggests that no change is warranted, then the current strategy of instruction is reinforced.

Hattie and Timperley (2007) suggest that classroom feedback flows in many directions. Teachers provide their input, peers offer assistance for each other, and parents provide support as well. The student is ultimately responsible for his or her own learning and should welcome corrective feedback from many sources. Evidence collected during the teaching of the lesson informs the teacher whether to adjust the content, pace, or strategy used during instruction (Popham, 2008). Questioning informs the teacher how well the students comprehend the material. Self-evaluation and reflection helps the students consider their own learning, thus providing feedback for themselves.

Evidence gained from student feedback helps drive the lesson forward. Wiliam (2010) stresses the point that feedback helps teachers answer the question, “What now?” For formative assessment to be effective feedback must be collected from students and the “What now?” question must be answered.

Heritage (2010) suggests six criteria to serve as a guide to learning while the student is
engaged in the learning tasks. To assist students in entering a learning task, the teacher must first present the learning target. The teacher would then guide the students through the learning cycle and the outcome would be the closing of the gap in understanding. This feedback loop, which is represented in Figure 1, would be considered part of short-cycle formative assessment.

Figure 1. The feedback loop. This figure illustrates short-cycle formative assessment. The learning target is presented, the feedback loop is engaged, and the gap in understanding is closed. Adapted from Heritage (2010), p. 11.

Instructional Correctives

Formative assessment is more than just collecting evidence through feedback; it includes adjusting instruction based on that evidence (Popham, 2008). During classroom instruction, just-in-time instructional corrections are critical to learners, and teachers must make quick decisions about which corrections might need to be made. Effective questioning provides teachers with evidence of learning and understanding. However, teachers must decide what to do with that evidence. Bloom’s mastery learning suggests that there must be some clear criteria of what
constitutes mastery in the classroom (Bloom, 1973). He believes teachers must have a specific idea about what their students are expected to learn and the teacher must help navigate a path for learning. The decisions teachers make must be consistent with the objective and learning goals suggested by the curriculum thus establishing instructional alignment (Guskey, 2007).

Defining AssessToday

The observation instrument, AssessToday, assesses a teacher’s use of short-cycle formative assessment during instruction. In order to design an observation instrument that would evaluate short-cycle formative assessment, two definitions of formative assessment were considered. The first was presented by Black and Wiliam (2009) and states:

Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited (p. 6).

The Council of Chief State School Officers (CCSSO) developed the second definition utilized. The committee defined formative assessment as “[a] process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students’ achievement of intended instructional outcomes” (McManus, 2008, p. 3).

These two definitions provided the framework for the constructs included in AssessToday. In sum, AssessToday needs to provide a platform for assessing the interactions between teachers and students. Based on the evidence collected, the instrument then needs to supply a method to assess the teacher’s ability to adjust his or her teaching strategy during the class period.

Constructs for AssessToday

Considering the definitions and the review of literature on formative assessment, seven constructs are included in AssessToday. The seven constructs are objectives, nature of
questioning, question quality, self-evaluation, observation of student affect, instructional adjustments, and evidence of learning. The developers used specific wording of the seven constructs in order to meet two important goals.

The first goal was to design an instrument that could be used by researchers or practitioners in any classroom situation. Since it is an observation instrument, the constructs were limited to actions observable during one period of instruction of approximately 45-90 minute duration. Although the instrument was designed to be used in any content area, the intent is for the observer to be a specialist in the content area they are observing. Certain aspects of the instrument require the observer to know the content in order to assess instructional adjustments or logical questioning. For this reason, the scope of this study is middle school and high school science and mathematics.

The second goal in the design of the instrument was to use wording that was familiar to both researchers and educators. For example, the phrase learning target is listed as essential in some of the research literature (Brookhart, Moss, & Long, 2008; Stiggins, 2009). Other phrases that are used in the literature to refer to learning targets include “learning intentions” (Wiliam, 2009, p. 37), “curricular aim” (Popham, 2008, p. 35), “learning goals” (Heritage, 2010, p. 37), and “learning expectations” (Wylie, Gullickson, Cummings, Egelson, Noakes, Norman, & Veeder, 2012, p. 22). Stiggins, Arter, Chappuis, and Chappuis (2006) point out that regardless of the label used, these are all statements of intended learning. Most educators are more familiar with the word objective and use it to refer to the intended learning for the day. The term objective was therefore selected for use in the instrument and refers to the learning target, or intended learning for the day’s lesson.

A clear description of each of the constructs and their indicators is described below.
Objectives. The first construct, objectives, was important for clearly communicating the purpose of the lesson, so that students know what they are learning. Objectives are much more than just written state or content standards, rather, they are the goals that drive the instruction and lead to curricular assessments during the class period (Brookhart, 1997). Many researchers prefer the term learning targets. Leahy et al. (2005) place clear learning targets as the first priority in the list of “nonnegotiable” strategies for effective classroom use of formative assessment (p. 20). Chappuis (2009) mentions that if teachers do not have clear and understandable learning targets, there is no way to design sound assessments. If teachers are going to use formative assessments to assist their students in the learning process, they must have a target in mind and know how they are going to achieve the target. Brookhart, Moss, and Long (2008) argue that clear learning targets give students ownership of their learning and increases a student’s motivation in the classroom. They also emphasize the importance of connecting classroom feedback to the established learning targets. This provides students with appropriate scaffolding as they progress towards the objectives of the lesson.

Shepard (2005) stresses the importance of learning objectives as students learn how to use self-assessments. She suggests teachers should provide students with rubrics that will inform students of the specific criteria expected of them throughout the learning process. This helps the students learn metacognition and to “internalize the standards” as they compare their work with the description and expectations written on the rubric (Shepard, 2005, p. 69).

Indicators for the objective construct. Heritage (2010) suggests that the process of formative assessment begins with identifying the learning targets. The targets, or objectives, are placed before the students as criteria for the day’s learning activities. If students are going to be cognizant of the objectives for the day, several strategies of communicating those targets should
be employed. Not only should teachers state the objectives for the day verbally but it is a good idea to write them on the board for students to see during instruction and for students to write them down also. Chappuis, Chappius, and Stiggins, (2009) stress the importance of students understanding the objectives of the day and what they need to do to achieve them. They suggest students write objectives in their own words, thus demonstrating a clear understanding of the day’s purpose. As teachers provide feedback during the lesson, students are reminded of where they are going and refocus their learning efforts on the intended objectives for the lesson. There are therefore four actions that can be observed during instruction that will indicate whether the objective construct has been thoroughly established:

1. Teacher stated the objective.
2. Teacher wrote the objective for students to see.
3. Students wrote down the objective.
4. Teacher revisited the objective throughout the lesson.

The objectives, therefore, are critical for the use of formative assessment in the classroom. They are the driving force of the day’s activities and one of the motivating factors for student learning. When teachers provide clear objectives, students are able to self-assess in such a way as to keep their learning moving forward. The objectives also provide a framework for feedback, which can be obtained through questioning and observation of student behaviors.

Questioning. Questioning is one of the most basic, yet most effective methods used by teachers to gain information, stimulate participation, and challenge thinking (Black & Harrison, 2001; Harlen, 2007; Korkmaz, 2009). Many researchers (Black & Harrison, 2001; Hannel, 2009; Redfield & Rousseau, 1981; Rowe, 1986; Taba, 1966; Tienken, Goldberg, & DiRocco, 2009) have studied the effects of questioning in the classroom and agree that questioning is a
crucial component in the teaching process and a vital part of formative assessment. Levin and Long (1981) indicate that teachers ask about 300 to 400 questions in any given day. Teachers ask questions to check for understanding, test knowledge of basic facts, diagnose student difficulties, and assist with classroom management. Wilen and Clegg (1986) state that questioning is “[t]he most influential single teaching act because of the power of the question to impact student thinking and learning” (p. 153). They indicate that there are two components at work in classroom discourse: the questions and the questioning. According to Black and Harrison (2001), the questions asked in the classroom are most effective when teachers are mindful of the quality of the questions and employ appropriate follow-up to students’ responses.

The quality of classroom questioning is an important component in classroom discourse. The quality of teacher’s questions can influence the degree to which students are able to recall prior knowledge and extend their thinking through scaffolding (Chin, 2006). Teacher questioning that elicits information about students’ understanding and encourages classroom discourse is an essential element in students’ learning of content. The use of open-ended questions and various levels of higher and lower order questions allow students opportunities to think and elaborate on their previous answers (Tienken, Goldberg, & DiRocco, 2009).

In addition to question quality, questioning structure is also important in affecting classroom discourse (Black & Harrison, 2001). Wilen and Clegg (1986), point out several effective questioning practices that can be used in the classroom. The questioning structure consists of wait time and various strategies for follow-up questioning.

In order to capture how teachers use questioning techniques to gain evidence of student understanding and achievement, the questioning construct is divided into two sub-categories, namely, question quality and nature of questioning. Question quality consists of three indicators:
open-ended questions, Bloom’s hierarchy of cognitive learning levels, and prior knowledge. Furthermore the nature of questioning consists of two indicators: wait time and follow-up questions. The researcher theoretically established these sub-levels and corresponding indicators through the literature review and practically reinforced them through personal experience and observation in the classroom.

Question quality.

Open-ended questioning. An important indicator of good questioning is the use of open-ended questions (Korkmaz, 2009). Open-ended questions seek creativity, encourage problem-solving, and may have multiple correct answers. They allow students the freedom to solve problems in new and innovative ways and discover unique problem solving strategies (Erdogan & Campbell, 2008). Additionally, class discussion is encouraged and students are provided opportunities to engage in higher-level reasoning, critical thinking, and evaluation. Almeida (2010) discusses the importance of using a dialogical approach to questioning (p. 589) which encourages dialogue between teacher and student and between students. This open-ended strategy encourages discussion where the teacher becomes a facilitator rather than a source for all questions to be asked and answered. Chin (2006) suggests that “students can be stretched mentally through sensitive teacher-led but not teacher-dominated discourse” (p. 1343).

Bloom’s high and low order questions. The second indicator of good questioning is based on an established hierarchy of cognitive levels as described by Bloom (1956). Although he was not writing about questioning, per se, his cognitive taxonomy does apply to the layered structure of questioning (Hannel, 2009). Redfield and Rousseau (1981) suggest that the low-cognitive-level questions correspond to the lower levels of Bloom’s taxonomy and the high-cognitive-level questions correspond to the higher levels. Shiman and Nash (1974) indicate
teachers do not necessarily need to follow a linear progression of Bloom’s cognitive levels but rather effective questioners move freely and efficiently between levels, both reinforcing facts and encouraging critical thinking and metacognition.

A meta-analysis conducted by Redfield and Rousseau (1981) indicated an overall effect size of .7292 for student achievement when teachers predominately used higher-order questions in the classroom. This positive effect on student achievement would mean that an average student who would score at the 50th percentile before the treatment could be expected to score at the 77th percentile after the treatment. The high-cognitive-level questions however, should be followed with a large number of low-cognitive-level questions to create a cognitive balance (Wilen & Clegg, 1986). Questioning structure also should be consistent with the objective of the lesson. If the objective for the lesson is to learn basic facts or learn a foundational content structure, then most questions would be low-cognitive-level questions dealing with memory and facts. If, however, the objective of the day is based on discovery or inquiry, the questioning structure would be completely different, consisting mainly of high-cognitive-level questions balanced with low-cognitive level questions.

Prior knowledge. For questioning to be effective, a students’ prior knowledge and the teachers scaffolding “grain size” must be consistently aligned (Popham, 2008, p. 33). Vygotsky (1962) addressed this issue in his discussion of ZPD. Teachers must challenge students to move from what they know to what they are capable of knowing. The grain size is crucial in completing the feedback loop that increases understanding and achievement (Sadler, 1989).

If students are going to build knowledge through the questioning process, they must have a foundation upon which to build (Korkmaz, 2009; Marzano, Pickering, & Pollock, 2001). This foundation is often referred to as prior knowledge and teachers need to appropriately activate
prior knowledge before the building of new knowledge can begin. Marzano, Pickering, and Pollock, (2001) suggest that teachers establish a “mental set” of available knowledge that students can use during the knowledge building process (p. 114). Students use their prior knowledge to organize, plan, and monitor most aspects of the learning process (Olvera & Walkup, 2010). The activation of prior knowledge also helps students with metacognition as they reflect on their current level of understanding. Prior knowledge is activated through open-ended questioning where students can check, refine, and enhance what they already know (Erdoşan & Campbell, 2008).

Nature of questioning.

Follow-up questioning. Good questioning skills in the classroom include follow-up questions. Multiple follow-up questions can diagnose specific areas of difficulty and give teachers a clearer idea of what correction is needed (Hannel, 2009; Shiman & Nash, 1974). Shiman and Nash (1974) discuss what they refer to as the “contextual questioner” (p. 250). The contextual questioner asks both factual and conceptual questions, moving effortlessly between the two, using multiple follow-up questions. During the follow-up questioning, the questioner looks for the moments of breakthrough and capitalizes on those learning opportunities. Hannel (2009) suggests that teachers should create a classroom environment that is conducive to questioning. Students should expect to be asked questions daily and the teacher should give every student the opportunity to justify and elaborate on the answers they give. Follow-up questions also focus students’ attention to specific areas the teacher wants to emphasize. Probing is another questioning strategy that encourages higher cognitive thinking and is effective to use when students are not accustomed to thinking or responding at higher cognitive levels. Students
may answer questions with low-cognitive responses, and one word answers, Follow-up questions help students to clarify, justify, and expand their initial responses (Wilen & Clegg, 1986).

*Wait time.* Researchers have also reported the importance of wait time during the questioning process (Almeida, 2010; Rowe 1974, 1986). The amount of time a teacher allows for student thinking and reflection is essential to the learning process. A study by Rowe (1974) found that teachers waited, on average, less than one second for a response. After that time, the teacher would either repeat the same question or call on a student for an answer. Rowe (1986) completed a follow-up study on teachers trained to wait 3-5 seconds before intervening. Student response improved dramatically, and therefore, there was much evidence of student learning and thinking. Wait time of this length is useful in discussion formats and gives greater confidence to students when they respond. Furthermore, Rowe listed ten positive effects on student learning and three positive effects noticed during teacher instruction. The positive effects on students included an increased length in their responses, greater participation, greater confidence, and fewer discipline problems. For teachers, the positive effects included a greater ability to be flexible in the classroom, an increased ability to ask appropriate questions, and higher expectations for student success.

*Self-assessment.* In the classroom it is important for students to become owners of their own learning (Leahy et al., 2005, Wiliam, 2008). This can be accomplished as students gain information about themselves through self-assessment (Black & Wiliam, 2009; Sadler, 1989). Andrade (2010) describes self-assessment as a process where students reflect on the quality of their own work, compare it with the stated objective, and make revisions as needed.

Teachers can improve self-assessment practices by encouraging students to reflect on their own learning through such techniques as “traffic lights” (Black et al., 2003, p. 22), (i.e. a
self-assessment technique where students display green, yellow, or red cards to indicate their level of understanding (Popham, 2008), a self-assessment inventory (Heritage, 2010), journaling (McMillan, 2010), and/or a reflection statement (Black et al., 2003).

Sadler (1989) suggested three necessary conditions for effective feedback: an understanding of the learning targets, a comparison between the current level of proficiency with the desired one, and engaging in appropriate action to move the learner forward. Hattie and Timperley (2007) worded these conditions in three simple questions that students could ask: “Where am I going? How am I going? and Where to next?” (p. 88). Teachers should encourage students to use self-evaluation to answer these critical questions during the instruction time. On the other hand, students should become self-regulated learners who automatically ask themselves these questions as they are in the process of learning. Through this process students learn to determine what they need to do to move their learning forward (Heritage, 2010).

Shepard (2000) suggests that self-assessment increases students’ responsibility for regulating their own learning and helps the relationship become more collaborative. Heritage (2007) agrees that students and teachers should develop a shared understanding of the student’s current development. This encourages students to take ownership for their learning and view the classroom as a safe place to take risks. Also, students who become more interested in classroom practices have longer attention spans and behave more appropriately in the classroom. Students who self-assess understand the criteria of achievement and are more likely to master the objectives of the lesson.

*Indicators for the self-assessment construct.* In sum, there are three main indicators for the self-assessment construct. The first indicator is that the teacher initiates self-evaluation strategies or tools during instruction. The second indicator is that the strategies or tools are
clearly defined and related to the objectives of the lesson. The third indicator is that students both understand how to assess their own learning and are able to use the strategies to assess their own learning. Students demonstrate continual use of self-evaluation strategies throughout the class period.

Student affect. Student affect includes students’ attitudes, interests, values, and dispositions (Popham, 2005). Teachers must be cognizant of how their students are receiving instruction (Cauley & McMillan, 2010). Much can be learned from students by observing their body language, facial expressions, and quality of their work. Stiggins (2010) points out that the successful use of formative assessment is not merely quality assessment and appropriate instructional adjustments, but also includes careful management of the emotional state of students. Popham (2005) suggests that student affect should be essential to teachers since student affect influences classroom behavior and engagement. He also supports using student affect as a component in any accountability strategy for education. Bell and Cowie (2001) suggest that formative assessment practice by teachers “is more than just an intellectual and professional activity; the feelings of teachers and students are centrally engaged” (p. 45).

Sensitivity to student affect in the classroom supports student’s motivation and self-esteem (Black & Wiliam, 2004). It also encourages students to take control of their own learning. Ownership of learning increases motivation, self-efficacy, and self-regulation (Brookhart, 2008; Miller & Lavin, 2007).

Heritage (2010) posits that what classrooms really need is a culture that is collaborative and supportive. She suggests shared responsibility between the teachers and all students, a classroom where risk taking occurs freely and frequently, and an environment that is characterized by mutual trust among teachers and students. Lave and Wenger (1991) refer to this
classroom culture as a community of practice and suggest students will work harder to become more competent members of the group when teachers and students have shared responsibilities. Student affect plays an important role in formative assessment. Achievement will increase when teachers effectively interpret affective signals and maximize opportunities for improvement.

**Indicators of the student affect construct.** There are four primary indicators that are observable in the classroom context and communicate information about student affect. The first indicator, student behavior, includes the degree of student responsiveness to teacher’s questioning, the amount of classroom engagement, indications of critical thinking, and display of confidence (Popham, 2005). The second indicator is the teacher’s behavior (Bell & Cowie, 2001). Teachers should continually be sensitive to student feelings, body language, facial expressions, and classwork (Leahy et al., 2005). The third indicator is the teacher’s focus. Teachers need to be learner-centered yet demonstrate balance between content and being sensitive to the affective needs of students (Shute, 2008). The final indicator is student interactions. A variety of student interactions are possible, which include student-to-student, student-to-group, and student-to-teacher interactions. Teachers should actively encourage student interaction and risk taking (Shepard, 2000).

**Instructional adjustments.** As discussed earlier in this paper, formative assessment can be effective in the classroom to inform teachers about adjustments they need to make in their instruction to facilitate better understanding. The key to instructional adjustments is that teachers use a different strategy for teaching than they used the first time the concept was taught (Guskey, 2007; Popham, 2008). As teachers collect evidence from their students during instruction, they analyze their own current instructional strategy and decide if it is working. If students are engaged in the learning and there is evidence of their understanding the content, the current
strategy is reinforced and the teacher can continue teaching with the current plan. If, however, students seem confused and are becoming frustrated with the material, the teacher should make some instructional adjustments.

Margolis and McCabe (2006) suggest the use of scaffolding to plan instructional adjustments in the classroom. If the task given to the students requires excessive effort on their part, it tends to make them feel inadequate and their motivation for learning decreases. Teachers should pay careful attention to student affect and adjust the “grain size” of instruction accordingly (Popham, 2008, p. 33). For example, many algebra students have difficulty solving word problems. When the teacher observes students becoming frustrated trying to find a solution, scaffolding in the form of sub-goals should be suggested that will guide students to successful completion of the task. The smaller grain size will help students stay motivated to continue working towards a solution. The effect this has on students is significant. Formative assessment’s emphasis on instructional adjustments and student affect increases motivation and encourages students to maintain high engagement in the classroom (Cauley & McMillan, 2010).

There are four primary components that need to be adjusted during instruction. They include the instructional strategy, instructional timing, instructional audience, and grouping strategies.

**Instructional strategies.** There are a number of methods and strategies available for teachers to collect feedback from students about their learning. Formative assessment is more than just collecting evidence: it includes adjusting instruction based on that evidence (Black & Wiliam, 2009). Wiliam (2010) have argued that formative assessment can inform a teacher that their strategy is better founded due to the evidence collected. Every teacher, however, needs to
correct his or her course(s) occasionally and formative assessment is one effective tool that can help teachers decide when adjustments need to be made.

Yin et al. (2008) proposes that in order for formative assessment to be successful, teachers must use the collected evidence from students to design new teaching strategies that will move students forward in their learning and close the gap. If the first teaching strategy did not effectually teach the intended concept, it logically follows that a different strategy should be tried the second time (Popham, 2008).

*Instructional timing.* The timing of the adjustment is especially important in the educational context. Guskey (2007) uses the term, “just-in-time corrections” to refer to those minor changes a teacher enacts to ensure minor problems in understanding do not become major problems later (Guskey, 2007, p. 13). At one extreme, a teacher may adjust their intended direction simply by noticing a frown on a student’s face and at the other extreme the teacher may wait until the student has had an opportunity to struggle with a difficult concept (Black & Wiliam, 1996). Both extremes are an appropriate use of the timing of formative assessment. The teacher has collected evidence from their students and made decisions about the best time to enact the change.

*Instructional audience.* The third instructional adjustment that may need to be made in the classroom is the instructional audience. The instructional audience is the group of individuals who are in need of the feedback. The feedback may be an explanation to an individual, a small group of students, or the entire class (Black & Wiliam, 2009). Evidence from classroom observations has shown teachers may interrupt the entire class to instruct one individual who has a question. On the other hand, a teacher might have several students with the same question and choose to answer each one individually. The problem with both of these
scenarios is the waste of instructional time. Time is precious in the classroom and every minute must be used to its fullest potential. The teacher must make wise adjustments based on the audience that needs the corrective feedback.

**Grouping strategies.** An overarching strategy that may be used in instructional adjustments is the grouping arrangements of students. According to Vygotsky (1962), the social component of learning is crucial in the development and construction of knowledge. When instructional adjustments need to be made in the classroom, various grouping strategies may be employed. If groups already exist in the classroom and adjustments need to be made, the teacher may redistribute students into different groups activating a new group dynamic. Davies (2009) provides several recommendations for establishing and using groups in the classroom setting.

One essential component to utilizing formative assessment, which teachers should keep in mind, is the idea and purpose of grouping. Teachers should intentionally construct the groups to maximize the advantages and minimize the disadvantages of group dynamics. Because of the immense diversity in the classroom, teachers need to carefully weigh how to arrange students in order to maximize learning. Each student should be given a specific task and responsibility in the group to ensure equity, improve motivation, and create ownership. The use of groups can be very effective during the feedback-corrective loop as students learn from each other.

**Evidence of learning.** The final construct of AssessToday is evidence of learning. Heritage (2010) stresses the importance of gathering quality feedback as evidence of student learning. She suggests four criteria for teachers to keep in mind when gathering evidence. First, the evidence should indicate that the student’s learning is aligned with the objectives of the day’s lesson. Second, sufficient evidence should be collected for teachers to make appropriate adjustments in teaching if needed. Third, teachers should gather evidence from a wide range of
understanding or skill levels in order to account for all students in the classroom. Finally, evidence should be gathered using more than one strategy to provide teachers with ample evidence that students have achieved the desired learning goals. Evidence of learning can come from several different sources during classroom instruction such as all-student responses, individual responses, and student artifacts.

*All-student responses.* One of the most common pieces of evidence used is an all-student response system (Heritage, 2010; Leahy, 2005; Wiliam et al., 2004). All-student responses are simply answers to a specific question that everyone in the room answers at the same time. Teachers can use several different methods to collect evidence of learning from the class as a whole. Wiliam et al. (2004) described the use of “traffic lights” as one technique, but many other techniques work as well. For example, teachers may use individual white boards, multiple choice cards, or electronic devices connected to a computer to have each student display an answer simultaneously.

*Individual responses.* Another common technique that can be used by teachers is individual responses (Black & Wiliam, 2009). Individual responses, like all student responses, require that every student give a response to a question. The key ingredient is that every student receives personal attention and feedback from the teacher. Andrade (2010) argues that individual responses are the best source of formative assessment feedback. When they are challenged to consider their own learning, it shifts the responsibility from the teacher onto themselves. This process helps students become self-regulated learners, increases their motivation for learning, and improves their self-efficacy (Stiggins, 2006).

*Student artifacts.* A third source of evidence for learning is student artifacts. Wiliam et al. (2004) discuss the importance of providing feedback on written work as well as providing
oral feedback. The comments made on written work seem to be most effective when grades are not associated with them because students tend to ignore the comments and place their emphasis on the grade instead (Butler, 1988). Popham (2008) strongly recommends leaving grades completely out of the formative assessment cycle. The function of formative assessment is to gather evidence of learning and adjust teaching strategies to enhance student achievement. Wiliam et al. (2004) also suggest providing feedback only and using the opportunity to increase learning. The artifacts can and should be used as instructional tools. The nature of the feedback is much more important than the amount of feedback. Complimentary phrases like, “nice job”, or critical statements like “wrong” do not provide any information that can move learning forward. Instead, suggestions for improvement are most effective. For example a teacher could ask, “can you explain this in more detail?”, or “what would happen if…?”.

Stiggins (2006) suggests engaging students in creating their own assessment activities similar to those that might be used as a summative score later. These written assessment artifacts can be traded with peers and used to evaluate learning. This type of activity encourages students to be owners of their own learning and providing opportunities for self and peer evaluation. Teachers can also provide useful feedback to help students ask each other valid questions and provide correct answers.

Content Specificity

Although AssessToday was designed for general use in educational settings, it will need to be validated for each major content area. The Standards for Educational and Psychological Testing (1999) suggest assessment measurements should be considered for various subpopulations. Since teaching mathematics and science differs substantially from teaching other content domains (McMillan, 2007), reliability and validity measures will need to be
verified separately for each major content area. The content areas considered in this study are science and mathematics. Further, the standards suggest that age groups should also be considered as subpopulations. Thus, the scope of this study encompasses middle school and high school science and mathematics.

Harlen (2007) states that a teacher’s content knowledge is reflected in the questions they ask, the classroom discussions that ensue, and comments teachers make regarding student artifacts. He posits that teachers of science and mathematics have special knowledge and skills in problem solving, communication, and reasoning. Only observers who have similar content knowledge in science and mathematics will be able to determine if a teacher is correctly handling the content domain. Heritage (2010) states that “effective formative assessment requires teachers to have the necessary depth of content knowledge to answer three essential questions: 1) Where are my students going? 2) Where are they now? 3) Where to next?” (p. 103). Observers using AssessToday need to identify if teachers are effectively meeting these requirements. For this reason, observers should be proficient in the content area they observe.

Significance and Conclusion

The discussion above describes the various constructs included in AssessToday and supports them with a thorough review of the literature. Many researchers and scholars have voiced the need for instruments that would help teachers improve their use of formative assessment in the classroom. AssessToday is such an instrument and if validated will fill a gap in formative assessment research.
The proposed research is significant for both researchers and educators.

For researchers:

- AssessToday will provide a baseline for the study of additional research in short-cycle formative assessment.
- The information gained from this validation study will be helpful in future studies by providing a thorough literature review on short-cycle formative assessment.
- AssessToday will also be the first of its kind to be used in determining how well teachers use short-cycle formative assessments. It was designed to be used across the curriculum and across grade levels; however, this study only validates AssessToday for mathematics and science at the middle and high school grade levels. Additional studies will need to be completed to ensure that AssessToday is valid and reliable in contexts other than mathematics and science and within grade levels prior to or following middle school and high school.
- The validation of AssessToday will also allow for longitudinal studies to ascertain whether it will provide guidance for teacher improvement over time.

For educators:

- AssessToday will serve as a powerful tool for administrators. They will have a tool that will provide information about how well teachers use short-cycle formative assessment. This information will inform administrators about possible professional development for the teachers in their schools or districts.
- AssessToday will instruct teachers on how to improve their use of formative assessment in the classroom.
• Professional learning communities (PLC) will have access to a useful protocol that will enhance team building and professionalism.

AssessToday was designed to be easily understood, comprehensive, and manageable. The language for this protocol is both clear and straightforward making it accessible to all teachers, administrators, or PLCs interested in developing their formative assessment skills. Since AssessToday is itself a formative assessment tool, teachers should feel comfortable using it in their classrooms. Formative assessment is a proven strategy for improving student achievement and AssessToday will be an effective tool in assisting teachers to use it to their fullest potential.
CHAPTER 3

METHODOLOGY

The literature review in the previous chapter described a process of teaching referred to as short-cycle formative assessment. Previous researchers demonstrated that the appropriate use of short-cycle formative assessment can have a positive impact on student achievement (Black & Wiliam, 1998; Leahy et al., 2005; Rodriguez, 2004; Wiliam & Thompson, 2007). Researchers have also shown that teachers have a difficult time integrating effective formative assessments into their teaching practices (Clark, 2011; Hargreaves, 2005; Leahy et al., 2005; McMillan et al. 2010; Robinson, Reed, & Strauss, 2011; Wiliam & Thompson, 2007). Cizek (2010) and Wiliam (2010) further mention the lack of available instruments available to evaluate a teacher’s use of formative assessment in the classroom. Therefore, there is a need to develop instruments that will help teachers improve their use of short-cycle formative assessment in the classroom (Cizek, 2010; Gallagher, 2010).

AssessToday was created to address these issues. Since AssessToday is a newly developed observation instrument, evidence must be provided to support its validity as a measure of short-cycle formative assessment and it must produce reliable results. Further, AssessToday must be established as valid and reliable in specific content domains and grade levels as recommended by the American Educational Research Association (AERA), the American Psychological Association (APA), & the National Council on Measurement (NCM), (AERA, APA, & NCM, 1999). The purpose of this study was to establish the validity and reliability of AssessToday in the content domains of science and mathematics for middle school and high school grade levels.
Research Questions

Four research questions guide this study:

1) Does AssessToday demonstrate content validity as determined by experts in:
   a. the field of formative assessment
   b. professionals in the field of science and mathematics education?

2) Does AssessToday demonstrate reliability in science and mathematics as determined from inter-rater reliability coefficients of trained observers of AssessToday?

3) Does exploratory factor analysis (EFA) identify any underlying factor structures in AssessToday?

4) Does AssessToday display evidence of construct validity as determined by content validity, the EFA, and reliability?

Scoring Procedure for AssessToday

AssessToday was designed in the form of a seven-by-four rubric where the seven core constructs of formative assessment were placed as rows and the four levels of efficiency were placed as columns (see Appendix A). A rubric was chosen for AssessToday due to the specificity that scoring rubrics can provide. Moskal and Leyden (2000) comment that rubrics address concerns of subjectivity and bias by formalizing scoring criteria. Stellmack et al. (2009) suggest well-developed rubrics can reduce rater discrepancy, improve inter-rater reliabilities, and improve consistency. Four levels of proficiency were chosen, rather than three or five, in order to force raters to choose a more definitive ranking. Observers tend to use the middle rankings when available instead of searching the evidence for distinguishing clues (Newell, Dahm, & Newell, 2002).
In addition to the AssessToday observation protocol, a handbook has been written that serves as a supplement to implement AssessToday. The handbook describes the various indicators for each construct in each level of short-cycle formative assessment (See Appendix B). The highest level for each construct would be scored as a three, the second highest a two, the third highest a one, and the lowest level a zero (see example in Table 2). Scoring the rubric in this manner provides scores that can be correlated using the percentage of rater agreement, Cohen’s kappa, and the Fleiss kappa (Smith, Vannest & Davis, 2011). This method of scoring is also in agreement of the recommendations of the Standards for Educational and Psychological Testing (AERA, APA, & NCM, 1999).

Table 2

Scoring Example for AssessToday.

<table>
<thead>
<tr>
<th>Core Constructs of Formative Assessment</th>
<th>Low Level (0 Points)</th>
<th>Moderately Low (1 Point)</th>
<th>Moderately High (2 points)</th>
<th>High-Level (3 Points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives (Clear Targets)</td>
<td>No objective is stated or written.</td>
<td>Teacher either states objective or has objective written for students to see.</td>
<td>Teacher states objective, has objective written for students to see, and students write objective for themselves.</td>
<td>Teacher states objective, has objective written for students to see, restates objective throughout the lesson, and students write objective for themselves.</td>
</tr>
</tbody>
</table>

Research Design for AssessToday Validation and Reliability

Two primary issues that must be resolved for any evaluation instrument to produce accurate results are validity and reliability (Bitner & Kratzner, 1995). The validity issue considers whether there is sufficient evidence relating the constructs to learning and theories of learning. Daniel and Siders (1994) suggest that validity is determined when an instrument measures what it is intended to measure. Validity is the most crucial criteria in the development

Content Validity

Content validity is defined by Haynes (2001) as “The degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose” (p. 75). Cronbach and Meehl (1955) suggest that content validity is established by demonstrating that the items from the instrument are a subset of the known information about the topic under investigation. Content validity is also considered a subjective measure of how appropriate the items seem to a group of reviewers who have some knowledge of the subject matter (Litwin, 1995). According to Bitner and Kratzner (1995), any instrument is considered to have content validity when it is connected to a body of research that is part of an accepted theoretical framework within the discipline.

The content validity of AssessToday, therefore, was assessed through a three-part process as suggested by the definition and descriptions above.

(1) A thorough review of the formative assessment literature was conducted in order to make sure the content of AssessToday was relevant to and representative of a teacher’s use of short-cycle formative assessment (Haynes, 2001).

(2) Based on the literature review, the most important elements of short-cycle formative assessment and their indicators were compared to the AssessToday rubric which served as the observation protocol. This process demonstrated the connection between AssessToday and the accepted theoretical framework found in the literature.

(3) Feedback from experts in the field of formative assessment was obtained for suggestions and improvements. Litwin (1995) makes the point that, “The assessment
of content validity typically involves an organized review of the survey’s contents to ensure that it includes everything it should and does not include anything it shouldn’t” (p. 35).

I collected feedback from two groups of experts. The first group of experts was 9 mathematics and science faculty members from 6 Texas Universities. All 9 experts were serving as project directors for the Texas Teacher Quality Grant program, were interested in the further development of formative assessment, and were skillfully trained in improving teacher quality in the classroom. I believed these experts had sufficient knowledge and insight into formative assessment to critically examine AssessToday and offer professional advice for improvement. The second group of experts was 5 of the foremost experts in the field of formative assessment. All five are researchers, authors, and well known for their work in the field of formative assessment.

The three part process will satisfy the afore mentioned criteria for content validity.

Reliability

Reliability is defined by Carmines and Zeller (1979) as “the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials” (p. 11). Brandon, Taum, Young, and Pottenger (2008) suggest that reliability assessment of observation instruments should include both consistency and consensus. Consistency is the degree to which various observers are consistent in their assignment of codes and consensus is the degree to which the observers agree with each other. Inter-rater reliability measures are essential in the evaluation of the quality of assessment instruments (Nichols, Wisner, Cripe, & Gulabchand, 2011).
Jonsson and Svingby (2006) reviewed 75 studies where assessment instruments were evaluated on a variety of criteria. More than half of the studies reviewed reported on inter-rater reliability and most of them, 32 out of 75, used the percentage of agreement between two raters as a measure. Researchers frequently use percentage of agreement due to its ease of use and validity for a variety of data types. Other statistics used to measure the degree of agreement between two raters included: The Pearson r, the Spearman rho, Cronbach’s alpha, and Cohen’s kappa. These statistics are applicable in specific contexts with various data types (Stellman, Konheim-Kalkstein, Manor, Massey, & Schmitz, 2009).

Fleiss kappa is a statistic that measures inter-rater reliabilities between more than two raters (Feiss, 1971; Nichols et al., 2011). An important component of Fleiss kappa is that it takes into account the amount of agreement expected due to chance. Thus, the kappa value provides an accurate value of the amount of consensus between a group of observers of three or more. The Fleiss kappa is particularly well-suited for ordinal scale data (Nichols et al., 2011).

The reliability for AssessToday was analyzed through correlational data from 36 item-by-item paired observations using two measures. The first measure analyzed the percentage of agreement between two observers, found by taking the sum of score agreements and dividing it by the sum of score agreements and disagreements and then multiplying by 100 (Baker, Gersten, Haayer, & Dingle, 2006; Kazdin, 1982). Using percentages alone can result in high reliability measures even if the raters assigned their scores randomly. Cohen’s kappa overcomes this issue by taking out agreements occurring by chance (Cohen, 1960). Cohen’s kappa was calculated using the formula $K = \frac{P(a)-P(e)}{1-P(e)}$, where $P(a)$ is the percentage of agreement and $P(e)$ is the expected percentage of agreement. Both the percentage of agreement and Cohen’s kappa provide accurate reliability measures and are used together in other studies of observation.
Fleiss kappa is a measure of inter-rater reliability that is used when three or more observers rate the same event (Fleiss, 1971). In order to examine the reliability of AssessToday among several observers, the Fleiss kappa was calculated from the scoring of recorded classroom videos by trained observers. The formula for calculating the Fleiss kappa is very similar to the formula used to calculate Cohen’s kappa: $k = \frac{p(o) – p(e)}{1 – p(e)}$. For the Fleiss kappa, p(o) is the observed proportion of agreement and p(e) is the expected proportion of agreement by chance (Sim & Wright, 2005).

Criterion-related reliability was also analyzed. Criterion-related reliability is determined when an observer’s score is compared to the score of an expert coder who has demonstrated consistent and objective scores. Frick and Semmel (1978) suggest that criterion-related reliability is a more useful measure when determining inter-rater reliability for observation instruments. They also propose criterion-related reliability will help support the generalizability of the observation instrument.

In order for AssessToday to be considered reliable, inter-rater and criterion-related correlation values should meet appropriate standards. The Standards for Educational and Psychological Testing (AERA, APA, & NCM, 1999) state that “the choice of estimation techniques and the minimum acceptable level for any index remains a matter of professional judgment” (p.31). In order to determine how many paired observations were needed to establish inter-rater reliability, I reviewed published documents concerning observation instruments that would provide information on an appropriate number. The following table (Table 3) records the
different observation instruments and the number of paired observations considered for inter-rater reliability associated with the corresponding researcher(s).

Table 3

*Researchers and the Associated Number of Observation Pairs*

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Observation Instrument</th>
<th>Number of pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appledoorn (2004)</td>
<td>Classroom Observation Protocol (COP)</td>
<td>6 (p. 67)</td>
</tr>
<tr>
<td>Applin (2005)</td>
<td>Culturally Responsive Teaching (CRT)</td>
<td>12 (p. 99)</td>
</tr>
<tr>
<td>O’Malley, Moran, Haidet, Seidler, Schneider, Morgan, Kelly, and Richards (2003)</td>
<td>Measuring Student Engagement in Health Professional Settings (STROBE)</td>
<td>7 (p. 98)</td>
</tr>
<tr>
<td>Marshall, Smart, and Horton (2009)</td>
<td>Electronic Quality of Inquiry Protocol (EQUIP)</td>
<td>16 (p. 308)</td>
</tr>
<tr>
<td>Wright and Craig (2011)</td>
<td>Tool for Assessing Responsibility-Based Education (TARE)</td>
<td>18 (p. 211)</td>
</tr>
</tbody>
</table>

Gwet (2010) discussed in detail the sample size needed for inter-rater reliability studies. Based on estimated values for agreement and error, the reliability study on AssessToday would require between 13 and 23 paired observations. Cantor (1977) suggested similar numbers. Based on the number of paired observations from the literature and professional opinions from statisticians, I decided to consider at least 16 paired observations for each content area: sixteen for mathematics and 16 for science. The large number of paired observations will provide sufficient evidence for the reliability coefficient of AssessToday in mathematics and science.

For the percentage of agreement between two raters a threshold of 80% or above is considered acceptable (Carmines & Zeller, 1979). Since Cohen’s kappa and the Fleiss kappa
remove chance agreements, the acceptable level has a slightly different interpretation as noted by Landis and Koch (1977):

1.00 - 0.81 – Almost perfect agreement
0.80 - 0.61 – Substantial agreement
0.60 - 0.41 – Moderate agreement
0.40 - 0.21 – Fair agreement
0.20 - 0.00 – Slight agreement.

If inter-rater or criterion-related reliability values fall below the kappa threshold of 0.70 as recommended by Landis and Koch (1977), I would recommend that the AssessToday constructs be re-evaluated to determine if changes need to be made to the wording or structure. Further, I would also recommend additional training for observers using the instrument.

Exploratory Factor Analysis

Henson and Roberts (2006) recommend conducting exploratory factor analysis (EFA) during instrument development in order to investigate the relationships between variables. They further suggest that EFA can be used to determine the theoretical constructs that underlie a given set of data and the extent to which they represent the original variables. Kieffer states that “the utilization of factor analytic techniques in the social sciences has been indelibly intertwined with developing theories and evaluating the construct validity of measures” (p. 75). In order to examine the internal structure and provide support for construct validity of AssessToday, EFA was conducted on the collected data using SPSS 20.

EFA is a complicated analytic technique in which the primary concern is to group a larger set of variables (constructs) into a smaller set (factors) based on the consistency of the data (Keiffer, 1999). Grouping the constructs together in this way reveals latent relationships that
might go unnoticed otherwise. Because of the mathematical rigor involved, computer programs are used to manipulate the data and extract the factors. There are many methods of factor extraction. I decided on principle component analysis (PCA) as it is one of the most popular (Henson & Roberts, 2006) and provides clear results for a variety of situations. I also considered evidence from principle axis factoring (PAF) to help interpret the meaning of the extracted factors.

The first step in finding the factors was to calculate a square correlation matrix from the data which correlates each construct to every other construct in the assessment instrument. Once the correlation matrix was created, the PCA method calculated an associated factor pattern matrix. The rows are composed of the constructs of the instrument and the columns are composed of the relative weights (loadings) of each construct to the extracted factor. The loadings are simply the Pearson $r$ correlations between the variable and the factor and thus will range from positive one to negative one. With the Pearson $r$, a coefficient of 0.78, for example, carries the same strength as a value of -0.78. For this reason the absolute value of the loading was considered when examining the extracted factor structure. The loadings were considered noteworthy when their magnitudes were greater than 0.30 (Stevens, 1999). The closer the magnitudes of the loadings are to one, the greater correlation there is between the construct and the factor.

Once the factor pattern matrix was calculated, associated eigenvalues were computed by summing the squared factor pattern loadings down the columns of the matrix. Eigenvalues thus represent the amount of variance associated with each factor and range from 0.0 to the total number of constructs in the analysis. For AssessToday, the eigenvalues would sum to seven since there are seven constructs and if the factor analysis calculated an eigenvalue of 3 for the
first factor, the amount of variance it describes would be three-sevenths (3/7) or 42.9% of the total variance. For this reason, the relative strength of the eigenvalue is determined in relationship to the number of constructs in the instrument.

For determining the number of factors to retain from the PCA, Stevens (2002) suggests three methods. The first method was to consider the factors with eigenvalues greater than one. This is the default setting in SPSS 20 and is thus used by many researchers; however, this is not an absolute rule in factor retention. Stevens also comments that this rule is somewhat arbitrary and should be used with other evidence including the scree test and the total amount of variance accounted for in the retained factors summing to at least 70% (the 70% rule).

A second method used to determine factor retention was the scree test, which is a graphical technique in which the factor numbers are listed along the X-axis and the magnitude of the eigenvalue is listed along the Y-axis. Stevens (2002) recommends retaining only those eigenvalues in the sharp descent before the graph begins to level off.

In addition to the eigenvalue greater than one rule and the scree test, Stevens (2002) also recommends considering the total variance accounted for in the factors. He suggests retaining enough factors to account for at least 70% of the variance, and in some cases up to 85% of the variance.

Ultimately, it is the responsibility of the researcher to examine the available evidence and retain the number of factors that is the best fit to the theoretical framework (Keiffer, 1999).

After the appropriate number of factors were retained, it was necessary to interpret the results. Keiffer (1999) posits that interpretation of the extracted factors is almost always aided by the rotation of the factor pattern matrix of the PCA. Several different rotations have been suggested by researchers; however the varimax orthogonal rotation is the most popular (Keiffer, 1999).
The orthogonal rotation of the factor matrix redistributes the variance and is said to "clean up the factors" and make them easier to interpret (p. 80). With the common variance redistributed, the strength of each construct relates directly to the factors. Keiffer mentions that the varimax rotation produces factors that have noteworthy loadings for a small number of constructs and near zero loadings for the others.

I conducted an EFA following the procedure described above. The PCA method was used to calculate the factor pattern matrix, evidence was collected for factor retention using the eigenvalue greater than one method, the scree test, and retained factors until at least 70% of the variance was accounted for. To gain insights into the interpretation of the retained factors, the factor pattern matrix was rotated orthogonally using varimax rotation and considered in light of the theoretical framework of formative assessment as discussed in the literature review. I also considered the rotated matrix from PAF to help with the interpretation of the factors.

Construct Validity

Construct validity is defined by Haynes (2001) as, “The evidence and rationales indicating the degree to which data from an assessment instrument measures the targeted construct” (p. 75). Litwin (1995) suggests that construct validity is the most valuable, yet most difficult way of assessing a survey instrument and is a measure of how meaningful and useful the instrument performs when in practical use. He also mentions that it can only be ascertained after the instrument has been used in multiple settings, over various populations, over a number of years. Haynes (2001) posits that construct validity is the sum of multiple validity indices such as content, convergent, internal, and external.

According to Carmines and Zeller (1979), establishing construct validation requires three steps: (a) a set of theoretical concepts and their interrelations must be specified; (b) a method to
measure the constructs proposed by the theory must be developed and their relationships 
examined; and (c) the empirical evidence must be interpreted in light of the theoretical 
relationships of the constructs. What is required therefore to examine construct validity is that 
one is able to state several theoretically derived hypotheses involving the particular concept and 
examine the proposed relationships. The researcher would find evidence that suggests construct 
validity due to the proposed theory behaving as predicted. Researchers suggest that construct 
validity can only be established after an instrument has produced consistent results in several 
settings, by a variety of individuals over a number of years, and in addition must include other 
Messick (1995) posits that any evidence that sheds light on the meaning of the instrument scores 
becomes part of construct validity.

Based on the above discussion, construct validity cannot be completely verified by the 
current study; however, the current study considered the content validity, criterion-based 
reliability, an EFA, and inter-rater reliability to determine if construct validity was supported and 
confirmed by the research results (Messick, 1995; Haynes, 2001).

Standards for Validity

In addition to meeting the criteria as described in the literature above, the current study 
also met the requirement for validity put forth by AERA, APA, and NCM as defined in the 
Standards for Educational and Psychological Testing (1999). Validity is “the degree to which all 
the accumulated evidence supports the intended interpretation of test scores for the proposed 
purpose” (p. 11). The Standards define five sources of validity that developers should address in 
order to insure sufficient evidence for validity to be demonstrated.
1. **Evidence based on test content** – “Test content refers to the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring” (AERA, APA, & NCM, 1999, p11). Evidence based on test content was collected through a panel of experts.

2. **Evidence based on response processes** – Since AssessToday is an observation instrument, this source of validity deals with the evaluation performance of the observers. “Relevant validity evidence includes the extent to which the processes of observers or judges are consistent with the intended interpretation of scores” (p. 13). Detailed descriptions of each construct at each of the four levels of proficiency have been provided to ensure sufficient evidence. The instrument and indicators for each level of proficiency were field tested and examined by a panel of experts. Inter-rater reliabilities were also examined to confirm this source of validity.

3. **Evidence based on internal structure** – “Analysis of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (p. 13). This source of evidence was confirmed through a thorough literature review.

4. **Evidence based on relations to other variables** – This source of validity is based on external variables such as other tests that measure the same constructs. It was intended for evidence to be collected for this source of validity through paired observations of AssessToday and the Electronic Quality of Inquiry Protocol (EQUIP) (Marshall, Smart, & Horton, 2009). However, the logistics of scheduling expert observers for both instruments to be administered during the same class period were not met.
5. Evidence based on consequences of testing – This source of evidence is about the intended or unintended consequences of test use. The intended consequence of AssessToday is to gather information about teachers that will help them be more mindful in their use of formative assessment in the classroom. The test is also designed to be used by individual teachers or professional learning communities to help teachers improve their use of the short-cycle formative assessment. The unintended consequence of this instrument is for administrators to use the observation protocol for summative assessment purposes.

Participants

In order for AssessToday to be verified and reliability tested, two groups of participants were selected. The first was a group of observers selected to investigate a teacher’s use of the short-cycle formative assessment by using AssessToday in classroom situations. The observers included 3 university faculty, 3 doctoral students, and 6 classroom teachers. All observers were proficient in either science or mathematics and only observed teachers in their content specialty. The second was a group of 56 in-service science and mathematics teachers that were observed by the first group. The 6 classroom teachers who conducted observations were also among the 56 in-service teachers who were observed. All in-service teachers were participants in a mathematics or science professional development program offered through a large university in Texas.

Research Sequence

The following sequence served as a guideline for the data collection process and for the verification of both the validity and reliability of AssessToday. A pilot study was also conducted following the same guidelines and was used to inform me of any validity or reliability issues.
prior to the primary study. Specific information pertaining to the pilot study follows the research sequence for the primary study provided below.

Step 1: Assess Content Validity
a. Compare the constructs of AssessToday to the available formative assessment literature.
b. Validate AssessToday by a panel of experts i.e. practitioners who use short-cycle formative assessment in their teaching, university professors who teach and mentor its use, and professionals who study and research its application.

Step 2: Assess Inter-rater and Criterion-related Reliabilities
a. Train Observers
b. Collect Data
c. Conduct Correlation Statistics

Step 3: Conduct EFA

Step 4: Assess Construct Validity
a. Content validity
b. Inter-rater reliability
c. Criterion-related reliability
d. Results from EFA on the collected data

Ethics Considerations

Although the research presents a minimum amount of risk to participants, an IRB was presented to the IRB ethics board and was approved.
Pilot Study

The pilot study was conducted at a middle school in the southern region of the United States. The purpose of the pilot study was to collect feedback data pertaining to the reliability and validity of AssessToday. Permission from the school was granted to observe students in classrooms of three mathematics teachers. These teachers agreed to participate in the data collection process, which included a training session, scoring AssessToday in paired observations, and allowing lessons to be observed by the other participants. The teachers were also asked to discuss the scores they assigned to fellow participants with me while also providing feedback about the AssessToday constructs. I used this feedback to recommend improvements to the AssessToday protocol.

Training of the Participants

A group training session with the participants was not possible due to scheduling conflicts. Consequently, individual training sessions on how to score AssessToday were conducted with each teacher during their planning periods. The participants were provided with the AssessToday protocol handbook and asked to read through the material prior to the training session. During the training session they received a detailed explanation of each of the seven constructs along with their associated indicators and a clear explanation of the scoring procedure. Once the participants felt comfortable with their ability to score AssessToday, a paired observation was conducted. During the observation, the observer and I recorded evidence pertaining to the constructs of AssessToday and individually rated the teacher. After the observation, the observer and I discussed each of the constructs and came to consensus on the score that seemed most appropriate.
Data Collection with Mr. Allen

Mr. Allen was the first teacher to work with AssessToday and conduct paired observations with me. He received his master’s in Administration and Educational Leadership. He had been teaching mathematics for 9 years, and for the past 6 years had been teaching seventh and eighth grade algebra. Mr. Allen was selected as teacher of the year in 2010.

Observation Results for Ms. Clark – Algebra I

After the training and initial observation, Mr. Allen and I observed Ms. Clark in an Algebra I class. She was teaching a lesson on solving linear equations. The class consisted of 9 males and 8 females. The students were seated in pairs with white boards and markers provided for each student.

After viewing the class period and scoring Ms. Clark using AssessToday, Mr. Allen and I compared scores and found that we agreed on 6 of the 7 constructs (Table 4). The construct we disagreed on was “Evidence of Learning”; Mr. Allen gave her a score of 2 and I gave her a score of 3. After considering the indicators as stated in the handbook, Mr. Allen agreed with the score given by me. Mr. Allen and I felt that Ms. Clark collected evidence from 90% of her students individually during the class period, used all student responses multiple times, and used an effective artifact as evidence of learning. Thus, a score of 3 seemed most appropriate.
Table 4

*The Scores for Ms. Clark (on a scale from 0 to 3)*

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Researcher</th>
<th>Mr. Allen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Question Quality</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Student Affect</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Observation Results for Ms. Reed – Algebra I

Mr. Allen and I also observed Ms. Reed in an Algebra I class. The class consisted of 8 males and 15 females. Students were sitting in groups of four with white boards and markers provided for each student.

After the observation, Mr. Allen and I assigned individual scores to Ms. Reed and compared them. On this observation Mr. Allen and I agreed on 5 of the 7 constructs (Table 5). The first construct we disagreed on was “Nature of Questioning,” which is described by two indicators: follow-up questioning and wait time. Mr. Allen and I noted the use of follow-up questioning several times; however, the teacher’s wait time was very sporadic, often answering her own questions immediately after asking them. I gave Ms. Reed a score of 1 because of her poor wait time and Mr. Allen gave her a score of 2 because of her frequent use of follow-up questioning. Since there are only two indicators for this construct each should weigh equally on
the score. After a brief discussion, I was persuaded by Mr. Allen that a 2 was a more appropriate score for this construct stating that her follow-up questioning was excellent, whereas her poor wait time was just below average.

A second construct that Mr. Allen and I disagreed on was “Observation of Student Affect”; Mr. Allen gave her a 2 and I gave her a 1. After a short discussion, Mr. Allen agreed that a score of 1 was more appropriate due to the absence of “risk taking,” which is an important indicator of the construct.

Table 5

*The Scores for Ms. Reed (on a scale from 0 to 3)*

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Researcher</th>
<th>Mr. Allen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Question Quality</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Student Affect</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Data Collection with Ms. Clark

The second teacher to work with AssessToday and conduct paired observations with me was Ms. Clark. She had taught mathematics for 4 years, and for the past 3 years had been teaching seventh and eighth grade algebra. Ms. Clark earned a BS in Mathematics Education
and a BS in Special Education. She was just a few credit hours shy of finishing her MS in school counseling at the time of the research.

Observation Results for Mr. Allen – AP Algebra I

After the training and initial observation, Ms. Clark and I observed Mr. Allen in an A.P. Algebra I class. The class consisted of 9 males and 14 females. Students were sitting in three curved rows with white boards and markers provided for each student. After the observation, Ms. Clark and I gave individual scores to Mr. Allen and compared them. On this observation we agreed on 5 of the 7 constructs (Table 6).

“Observation of student affect” was one of the constructs which we disagreed upon. I scored Mr. Allen a 2 and Ms. Clark scored him a 3. I pointed out to Ms. Clark that the indicators for the student affect construct at high-level (3) uses descriptive words like “continually,” “superb,” and being sensitive to “all” students, whereas the indicators for moderately-high (2) uses descriptive works like “generally,” “makes an effort to be,” and being sensitive to “most” students. After some discussion, Ms. Clark agreed that a score of 2 was more appropriate for this construct based on the wording of the indicators. A second construct we disagreed on was “instructional adjustments.” I gave Mr. Allen a score of 1, and Ms. Clark a score of 2. During the discussion of this construct, Ms. Clark commented that observing instructional adjustments were difficult during a lesson cycle. She felt that Mr. Allen did not need to make adjustments to his instruction and therefore should not be given a low score just because his lesson went well. I agreed with her in principle, but suggested the observer could only score what they see during the classroom period. I made note of the comment made by Ms. Clark as a possible addition to the AssessToday training in the future.
Observation Results for Ms. Reed – Algebra I

Ms. Clark and I also observed Ms. Reed in an Algebra I class. The class consisted of 8 males and 10 females. Students were seated in groups of various sizes ranging from one to four. Ms. Reed was teaching a lesson on how to solve linear inequalities. After the observation, Ms. Clark and I individually rated her and compared scores. On this observation we agreed on 7 of the 7 constructs (Table 7).
Table 7

*The Scores for Ms. Reed (on a scale from 0 to 3)*

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Researcher</th>
<th>Ms. Clark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Question Quality</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Student Affect</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Data Collection with Ms. Reed

The third teacher to work with AssessToday and conduct paired observations with me was Ms. Reed. She had been teaching mathematics for 4 years, the past 3 of which had been seventh and eighth grade algebra. Ms. Reed earned a BS in mathematics education and a Math Specialist degree.

Observation Results for Mr. Allen – AP Algebra I.

Ms. Reed and I observed Mr. Allen in an A.P. Algebra I class. The class consisted of 10 males and 6 females. Students were sitting in three curved rows with white boards and markers provided for each of them on their desks. Mr. Allen was teaching a class on central tendencies which covered mean, median, mode, and range. After the observation, Ms. Reed and I individually assigned scores to Mr. Allen and compared them. On this observation we agreed on 5 of the 7 constructs (Table 8). The constructs we disagreed on were “question quality” and
“self-evaluation”. Due to time constraints, Ms. Reed and I were unable to discuss the differences in our scores.

Table 8

The Scores for Mr. Allen (on a scale from 0 to 3)

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Researcher</th>
<th>Ms. Reed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Question Quality</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Student Affect</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Observation Results for Ms. Clark Algebra I.

Ms. Reed and I also observed Ms. Clark in an Algebra I class. The class consisted of 7 males and 10 females. Students were sitting in straight rows with white boards and markers provided for each student. Ms. Clark was teaching a discovery lesson on the slope of lines.

After the observation, Ms. Reed and I individually rated Ms. Clark and compared scores. On this observation we agreed on 6 of the 7 constructs (Table 9).

The construct we disagreed on was “instructional adjustments.” I scored Ms. Clark a 1 and Ms. Reed scored her a 2. During the discussion about the “instructional adjustments” construct, Ms. Reed stated that Ms. Clark adjusted her class strategy before the lesson began based on feedback she received from her previous class period. I pointed out that even though
instructional adjustment can occur between classes, the AssessToday protocol is based on observable actions that take place during a single class period.

Table 9

*The Scores for Ms. Clark (on a scale from 0 to 3)*

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Researcher</th>
<th>Ms. Reed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Question Quality</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Student Affect</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Inter-Rater Reliability for the Pilot Study

In considering inter-rater reliability, this study considered two measures: the percentage of agreement between two observers and Cohen’s kappa. I calculated the percentage of agreement using the formula $A = 100\frac{\text{agreements}}{\text{agreements} + \text{disagreements}}$ and Cohen’s kappa using the formula $K = \frac{P(a)-P(e)}{1-P(e)}$, where $P(a)$ is the percentage of agreement and $P(e)$ is the expected percentage of agreement. A summary of the results is found in table 10. Although three of the observations resulted in scores slightly below the recommended thresholds, the overall study yielded results within the accepted range.
Table 10

*Observation Scores from the Pilot Study*

<table>
<thead>
<tr>
<th>Observers</th>
<th>Observed</th>
<th>Agreements</th>
<th>Disagreements</th>
<th>Percent Agreement</th>
<th>Cohen’s Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher and Mr. Allen</td>
<td>Mrs. Clark</td>
<td>6</td>
<td>1</td>
<td>85.7</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Ms. Reed</td>
<td>5</td>
<td>2</td>
<td>71.4</td>
<td>0.62</td>
</tr>
<tr>
<td>Researcher and Ms. Clark</td>
<td>Mr. Allen</td>
<td>5</td>
<td>2</td>
<td>71.4</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>Ms. Reed</td>
<td>7</td>
<td>0</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>Researcher and Ms. Reed</td>
<td>Mr. Allen</td>
<td>5</td>
<td>2</td>
<td>71.4</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>Ms. Clark</td>
<td>6</td>
<td>1</td>
<td>85.7</td>
<td>0.81</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>34</td>
<td>8</td>
<td>81</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Summary of Results

AssessToday produced satisfactory results and proved to be sufficiently reliable on a small scale with personal training. The percentage of agreement of 81% and the Cohen’s kappa of 0.75, as indicated in Table 10, both fall within the acceptable standards for reliability of 80% for percentage of agreement and k = 0.70 for Cohen’s kappa. The Fleiss kappa was not calculated due to the small number of observers in the pilot study. To produce these results, I took the following steps for training observers: (1) individual reading of the AssessToday handbook, (2) a one hour training session on how to assign scores using AssessToday, and (3) a discussion of one non-graded paired observation with each participant.
Conclusions from the Pilot Study

I conducted the pilot study in order to collect information that would be beneficial in determining the validity and reliability of AssessToday. The feedback received from the participants provided support for the content validity of AssessToday. All three of the participants provided very positive feedback about the instrument. Ms. Reed mentioned to me that when she was being observed, AssessToday did not make her feel uneasy. She also mentioned that AssessToday provided a fair assessment of her class period and it would help her become more effective and also implement formative assessment into her classroom. All three teachers stated discontentment with the instrument currently used in their school to assess their teaching and would rather be assessed by an instrument like AssessToday.

One interesting observation occurred when the observer and I disagreed on a construct score. I noticed that 7 out of 8 disagreements, or 88% of the time, the observer awarded a higher score than I did. One possible explanation for this is rater bias. Since the three teachers who participated in the pilot study work together, have known each other for several years, and see each other daily, it is possible they scored each other higher than they would have scored an unknown teacher. Rater bias is discussed in the literature (AERA, APA, & NCM, 1999; Chatterji, 2003) as one possible fairness issue with assessments in general. However, in the case of AssessToday, the rater bias was influenced by collegiality and not by the wording of the constructs. I would recommend that rater bias be included as part of the training procedures for AssessToday.

A second observation made was the even distribution of disagreements on construct scores. The participating observers and I disagreed 8 times during the pilot study and 6 of the 7 constructs were disagreed upon at least one time. Student affect and instructional adjustments
were disagreed upon two times with objectives being agreed upon every time. Since some variation in observer scores is to be expected, these results do not indicate an obvious problem with any of the constructs or their indicators. The disagreements can be explained by the difference in scorer interpretation of classroom events or the difference in understanding of the constructs. I would recommend conducting more thorough training to help alleviate disagreements among observers.

Overall, the evidence from the pilot study supports the content validity of AssessToday. The seven constructs included in the instrument are observable and measureable in the mathematics classroom. The inter-rater reliability scores indicate that the terminology used to describe the levels of proficiency in AssessToday is clear and appropriate. It was also noted that the observers scored the teachers using all four levels of proficiency, 0,1,2 and 3. This is significant as it indicates all levels of proficiency are possible in classroom settings, thus giving support for the four-level structure of the instrument.

Conclusion

Recently, there has been much discussion about formative assessment in educational circles. Researchers have conducted several studies and subsequent articles have been written over this topic. There is, however, much left to do. Many hope that the use of formative assessment can bring about significant improvement in achievement and close the gap between the high and low achievers. For that to happen, formative assessment must be understood in the classroom context and its application must be mastered by teachers. Professional development programs must be put into place which will help ease teacher’ apprehensions about formative assessment. Teachers need a guide to help them along the path to greater efficiency and productivity in their use of formative assessment. AssessToday was designed for that purpose.
There is a tremendous need in education for tools and strategies that can have a positive effect on teachers and students. Short-cycle formative assessment is such a tool when used appropriately. The following chapters will describe the results from determining the validity and reliability of AssessToday.
CHAPTER 4

RESULTS

The purpose of this study was to determine the validity and reliability of AssessToday, a short-cycle formative assessment observation instrument. Data was collected and analyzed to determine content validity, reliability, and construct validity as described in Chapter 3. Chapter 4 will address the results for each research question. Chapter 5 will provide a discussion and conclusion of the findings.

The four research questions that guided this study were:

1) Does AssessToday demonstrate content validity as determined by experts in:
   a. the field of formative assessment
   b. professionals in the field of science and mathematics education?
2) Does AssessToday demonstrate reliability in science and mathematics as determined from inter-rater reliability coefficients of trained observers of AssessToday?
3) Does exploratory factor analysis (EFA) identify any underlying factor structures in AssessToday?
4) Does AssessToday display evidence of construct validity as determined by content validity, the EFA, and reliability?

Research Question 1

Does AssessToday demonstrate content validity as determined by experts in (a) the field of formative assessment and (b) professionals in the field of science and mathematics education?

I discussed three steps for data collection in Chapter 3 as a method to determine the content validity of AssessToday. The first step consisted of a thorough review of the formative assessment literature in order to make sure the content of AssessToday was relevant to and

The second step in determining the content validity of AssessToday consisted of comparing the constructs of AssessToday with the most important elements found in the literature. These include objectives, questioning, self-evaluation, student affect, instructional adjustments, and evidence of learning. Table 11 associates each of the seven constructs with references from the supportive literature. This step also satisfied the third source of validity, evidence based on internal structure, as indicated in the Standards for Educational and Psychological Testing (1999). A detailed explanation of the associated indicators can be found in Appendix B.

The third step in determining content validity was to receive feedback from various experts in the field of education for suggestions and improvements and to confirm that AssessToday was comprehensive in its approach to short-cycle formative assessment and not lacking in any vital components (AERA, APA, & NCM, 1999; Litwig, 1995). AssessToday was first sent to Dr. Jim Morrow, Regents Professor at the University of North Texas, who provided useful insights into the wording and structure of the constructs. Based upon his recommendations, the instrument was revised, adding clarity and consistency to the protocol. Dr. Morrow also recommended the development of a training manual for the users of AssessToday. The training manual can be found in Appendix B.
<table>
<thead>
<tr>
<th>Constructs</th>
<th>Supporting Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>Black &amp; Wiliam (2009), Bloom (1956), Brookhart, Moss, &amp; Long (2008), Chappuis (2009), Heritage (2010), Leahy et al. (2005), Popham (2008), Shepard (2005), Stiggins (2010)</td>
</tr>
</tbody>
</table>
AssessToday was also presented to a panel of experts in mathematics and science education for purposes of gaining feedback and suggested improvements. The panel of experts met together on two different occasions to discuss the wording and structure of the instrument. The panel consisted of 3 mathematics and 6 science content specialists who were also university faculty, educational researchers, and who have years of experience with teachers. Also present at the meeting were the two developers of AssessToday, one who is a mathematics content specialist and the other who is a science content specialist. Based on the feedback from the panel of experts, AssessToday was revised. The revisions to the constructs are noted in Table 12.

The revised instrument was reviewed a second time by the nine panel members in an effort to gain additional feedback and recommendations. No additional feedback or revisions were received from the expert panel.
## Suggested Improvements to the AssessToday

<table>
<thead>
<tr>
<th>Construct</th>
<th>Noted problems and recommendations</th>
<th>Action Taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>It was noted that the four levels were not well defined. There were various situations in the classroom that would not provide a clear level.</td>
<td>The criteria for the four levels were re-written for clarity and consistency.</td>
</tr>
<tr>
<td>Questioning</td>
<td>Two problems were noted during this discussion. (1) The high-level indicator for Bloom’s levels says “teacher used high-level questioning throughout the lesson”, however sometimes that is inappropriate. Wording should be changed to “an appropriate amount” of high level questioning. (2) What would happen if a teacher has mixed levels of proficiency for the five indicators? How would one rank them?</td>
<td>(1) Wording for Bloom’s levels was revised. (2) This issue was resolved by splitting the questioning construct into two parts, Question Quality and Nature of Questioning.</td>
</tr>
<tr>
<td>Self-Evaluation</td>
<td>No problems were noted</td>
<td>None taken</td>
</tr>
<tr>
<td>Observation of Student Affect</td>
<td>It was noted that the word “creative” under “student interactions” would be difficult to observe. There was also incorrect wording under “moderately-high teacher focus”.</td>
<td>The word “creative” was removed. The “moderately high teacher focus” was reworded for clarity.</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>The phrase “Just-in-time” was unclear to some.</td>
<td>A clear definition of “just-in-time” was added to the end of this construct.</td>
</tr>
<tr>
<td>Evidence of learning</td>
<td>Can an instructional adjustment be pre-planned and still satisfy this construct?</td>
<td>No changes were made to the indicator, however additional information was included in the handbook used during training.</td>
</tr>
<tr>
<td></td>
<td>Classroom grouping procedures would be difficult to observe.</td>
<td>This indicator was reworded for clarity.</td>
</tr>
<tr>
<td></td>
<td>The phrasing “group Response” was very confusing.</td>
<td>This indicator was reworded.</td>
</tr>
<tr>
<td></td>
<td>The word “verbal” under “moderately-high individual responses” was problematic.</td>
<td>The word verbal was removed</td>
</tr>
<tr>
<td></td>
<td>What is an authentic artifact? Why is it necessary to demonstrate learning?</td>
<td>The word authentic was removed.</td>
</tr>
</tbody>
</table>
Furthermore, AssessToday was sent out for review by seven of the foremost experts in the field of formative assessment. I received feedback from five of the seven experts: Dr. Susan Brookhart - consultant in educational assessment at Brookhart Enterprises LLC, Dr. Margaret Heritage - Assistant Director for Professional Development at the National Center for Research on Evaluation, Standards and Student Testing (CRESST) at UCLA, Dr. James McMillan – Professor of education, Virginia Commonwealth University, Dr. Jim Popham - Professor emeritus at UCLA, and Dr. Rick Stiggins - Exec Director, Assessment Training Institute.

The request was three fold:

1) Question 1: Does AssessToday appear to cover all the major constructs of short-cycle formative assessment that could be measured in a single classroom observation?

2) Question 2: Are all the constructs included in AssessToday valid constructs of short-cycle formative assessment that could be observed in a single classroom observation?

3) Question 3: What suggestions do you have to improve this instrument?

For questions one and two, there was general agreement that AssessToday covered all the major constructs of short-cycle formative assessment and that each could be measured in a single classroom observation. Stiggins stated, “The constructs to be evaluated in your framework are the appropriate ones for sure—good job; they are the right things to track and to help teachers develop” (R. Stiggins, Personal correspondence, June 2013). McMillan agreed, “Overall, I think this is a much needed type of instrument and has great promise to be able to provide more
systematic data on the extent to which the kind of formative assessment that really makes a
difference in learning” (J. McMillan, Personal correspondence, June 2013).

For question three, the primary suggestion to improve AssessToday was a refinement of
the wording of the construct indicators in order to add clarity and consistency. Brookhart
recommended changing the term objective to the phrase learning target since achieving a
learning target is more descriptive of what should be happening in the classroom. Stiggins,
Brookhart, and Heritage suggested the wording of the Objective construct needed the most work.
In general, their concern was not with the indicators per se, but that that the indicators were
missing a level of quality. They also suggested rewording the evidence of learning indicators;
suggesting that a level of quality was missing from the construct indicators (S. Brookhart, M.
Heritage, & R. Stiggins, personal correspondence, June 2013). The developers of AssessToday
edited the construct indicators as recommended by the experts. The term objective was changed
to learning targets and the indicators were reworded to add clarity and consistency to the
instrument.

The feedback received from the foremost experts in the field of formative assessment
provided support for the content validity of AssessToday. Although experts suggested
modifying the construct indicators, they agreed that AssessToday included core criteria
associated with short-cycle formative assessment.

In addition to feedback from experts in the field, AssessToday was extensively field
tested in various mathematics and science classrooms in Texas and Oklahoma by Dr. Colleen
Eddy, Dr. Pamela Harrell, and others. The field testing provided important information about the
instruments’ usability in classroom settings as well as additional feedback pertaining to the
appropriateness of the constructs and the content validity of AssessToday. No major revisions
were warranted, and I pilot tested the AssessToday Observation Protocol. The results from the pilot study were very positive and were discussed in Chapter 3.

Evidence of Content Validity from Observation Results

In order to further determine content validity, 12 observers were trained to score teachers using the AssessToday instrument. During the training, the observers read through the training material and discussed the meaning of each construct and associated indicators. The meaning for each level of proficiency was also addressed. Short video clips were viewed to demonstrate each construct and were made available for online review. The training lasted approximately two hours and was followed by a video recorded lesson of Mr. Green teaching a mathematics lesson. I had video recorded a mathematics lesson taught by Mr. Green and provided it to the observers for scoring. Observers assigned individual scores to the recorded lesson using AssessToday and then discussed the scores with me. In order to get a baseline measure of reliability for the group, I calculated a Fleiss kappa for the viewing of the Mr. Green video. The baseline kappa was .26 which falls in the fair agreement category for the Fleiss kappa coefficient (Landis & Koch, 1977).

When the training was complete, the observers were assigned teachers to score using the AssessToday protocol. When possible, teacher observations were viewed through the use of Facetime® using iPads®. Using Facetime® allowed observers the ability to view a single classroom in real time without the time and cost restraints of travel. Using this technology also provided opportunities for paired observations and proved to be an effective non-intrusive method to collect observation data.

Content validity was supported from the results of the trained observers who conducted 98 classroom observations of 56 participants (Appendix C and D). Participating mathematics
and science teachers were scored during a single class period by trained observers using AssessToday. Participants could receive a possible score of 0, 1, 2, or 3 on each of the seven constructs. The observers provided rationale for the scores using the descriptive indicators found in the AssessToday handbook. A summary of mathematics results are provided in Table 13 and illustrated in Figure 2. Most mathematics teachers scored in the middle two levels of proficiency for each of the seven constructs with the highest number in the moderately low level.

Table 13

*Frequency Table in Percentages (n=64 mathematics observations)*

<table>
<thead>
<tr>
<th>Construct</th>
<th>Low Level</th>
<th>Moderately Low Level</th>
<th>Moderately High Level</th>
<th>High Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>24%</td>
<td>61%</td>
<td>9%</td>
<td>6%</td>
</tr>
<tr>
<td>Question Quality</td>
<td>1%</td>
<td>42%</td>
<td>50%</td>
<td>7%</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>1%</td>
<td>44%</td>
<td>46%</td>
<td>7%</td>
</tr>
<tr>
<td>Self-Evaluation</td>
<td>15%</td>
<td>58%</td>
<td>25%</td>
<td>0%</td>
</tr>
<tr>
<td>Student Affect</td>
<td>3%</td>
<td>34%</td>
<td>54%</td>
<td>9%</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>17%</td>
<td>59%</td>
<td>23%</td>
<td>1%</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>1%</td>
<td>44%</td>
<td>49%</td>
<td>6%</td>
</tr>
</tbody>
</table>
Figure 2. Assess Today construct distribution by percentage for 64 mathematics teacher observations. The highest score for formative assessment proficiency yielded the lowest percentage of teachers.

A summary of science results are provided in Table 14 and illustrated in Figure 3. As with the mathematics observations, most science teachers scored in the middle two levels of proficiency for each of the seven constructs with the highest number in the moderately low level. A brief discussion of each construct for both mathematics and science follows Figure 3.

Table 14

Frequency Table in Percentages (n=34 science observations)

<table>
<thead>
<tr>
<th>Construct</th>
<th>Low Level</th>
<th>Moderately Low Level</th>
<th>Moderately High Level</th>
<th>High Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>20%</td>
<td>36%</td>
<td>36%</td>
<td>8%</td>
</tr>
<tr>
<td>Question Quality</td>
<td>5%</td>
<td>50%</td>
<td>38%</td>
<td>7%</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>11%</td>
<td>47%</td>
<td>27%</td>
<td>15%</td>
</tr>
<tr>
<td>Self-Evaluation</td>
<td>15%</td>
<td>52%</td>
<td>26%</td>
<td>7%</td>
</tr>
</tbody>
</table>
As with the mathematics teachers, the highest score for formative assessment proficiency yielded the lowest percentage of teachers. Most teachers were scored at a moderately low level of formative assessment competency.

Objective construct for mathematics and science. The literature review suggested formative assessment is most effective when the objectives are stated and written by the teacher, written down by the students, and revisited frequently during instruction. Figure 4 indicates that most of the observed teachers are weak in their communication of the objectives. It also suggests that science teachers are somewhat more proficient than mathematics teachers.
Figure 4. Objectives construct by percentage of scores for mathematics and science. The highest level of proficiency for the objectives construct yielded the lowest number of teachers for both science and mathematics.

Question quality construct for mathematics and science. Question quality consists of the use of open ended questions, a mixture of high and low level questions consistent with Bloom’s taxonomy, and connection to students’ prior knowledge. Figure 5 indicates that science and mathematics teachers are very similar in the quality of the questions they ask in the classroom. This construct had the smallest percentage of teachers in the lowest proficiency level.

Figure 5. Question quality construct by percentage of scores for mathematics and science.
Nature of questioning construct for mathematics. The nature of questioning pertains to the procedural aspects of questioning; wait time and follow-up questions. Figure 6 illustrates that most teachers were scored in the middle two proficiency levels. Mathematics teachers were scored somewhat higher overall, however more science teachers were scored in the high level for this construct.

Figure 6. Nature of questioning construct by percentage of scores for mathematics and science. Only 1% of mathematics teachers were scored at the lowest level for this construct.

Self-evaluation construct for mathematics and science. The self-evaluation construct consists of both teachers’ observed behaviors and the observed behaviors of students. Teachers should initiate a variety of self-evaluation strategies for students to employ, and students should understand how to engage the strategies for self-learning. Figure 7 indicates that none of the mathematics teachers observed were placed at the highest level for this construct and only 5% of the science teachers. Over 70% of the teachers scored in the lowest two level of proficiency for this construct.
Figure 7. Self-evaluation construct by percentage of scores for mathematics and science.

Student affect construct for mathematics and science. Four indicators describe the student affect construct: student behavior, teacher behavior, teacher focus, and student interactions. Figure 8 indicates that both mathematics and science teachers received good scores. Overall, the scores for this construct were the highest among all the constructs for both mathematics and science teachers. Over half of the teachers observed scored in the top two levels of proficiency.

Figure 8. Student affect construct by percentage of scores for mathematics and science.

Instructional adjustment construct for mathematics. The instructional adjustment construct consists of four indicators: instructional strategies, instructional timing, instructional
audience, and grouping strategies. Figure 9 indicates that most of the mathematics and science teachers scored similarly and overall in the lower ranges of competency.

Figure 9. Instructional adjustments by percentage of scores for mathematics and science. The scores for mathematics and science teachers were closely aligned for this construct.

Evidence of learning construct for mathematics and science. Teachers obtain evidence of learning from three sources: responses from all students in group answers, responses from individual students, and artifacts of learning that students produce. Figure 10 illustrates the similarity between the scores for mathematics and science teachers.

Figure 10. Evidence of learning construct by percentage of scores for mathematics and science.
The comparison between the scores for the mathematics and science teachers, as shown in Figures 4 through 10, indicates several similarities between the two content areas. First of all, most of the teachers scored in one of the two middle categories of proficiency with the highest percentage of scores falling in the moderately low level. The objective construct received the lowest overall percentage score for both mathematics and science teachers. In addition, the student affect construct received the highest overall percentage score for both groups.

Evidence for Content Validity

The distribution of the scores from the 98 teacher observations provide support for content validity of the AssessToday instrument. Chatterji (2003) discusses the importance of experimental results reflecting the construct they are intended to measure. One notable fact is that every level of proficiency is used for one or more of the observations, thus the results reflect the confines of the constructs. This indicates that the proficiency levels are appropriately defined. Also interesting to notice is that the overall distribution for both mathematics and science scores resembles a normal distribution. Both of these results meet the conditions for content validity set forth by the Standards for Educational and Psychological Testing (AERA, APA, & NCM, 1999).

Research Question 2

Does AssessToday demonstrate reliability in science and mathematics as determined from inter-rater reliability coefficients of trained observers of AssessToday?

Reliability is an important characteristic of any protocol, especially those used for educational assessment (Bresciani, Oakleaf, Kolkhorst, Nebeker, Barlow, Duncan, & Kickmott, 2009). I examined inter-rater and criterion related reliabilities using the percentage of agreement
between raters and Cohen’s kappa. Brandon et al., (2008) commented that data cannot be regarded as valid unless they are also reliable.

The inter-rater and criterion-related reliability coefficients were determined from 36 paired classroom observations, 16 for mathematics and 20 for science. The paired observations were conducted by 11 of the 12 trained observers of AssessToday. The only condition for the pairing was the observer’s content specialty and availability to participate in a paired observation. One trained observer did not conduct any paired observations. Two of the observers, one in mathematics and one in science, were designated as the experts for the criterion-related reliability. The two experts were content specialists who had several months of experience with the instrument, sufficient understanding of the constructs, and demonstrated consistent scores of teachers through multiple observations.

A summary of the inter-rater reliability results for mathematics and science is presented in Table 15, and a summary of the criterion-related reliability for mathematics and science is presented in Table 16. The average inter-rater percentage of agreement for mathematics was 81% with a corresponding average kappa value of 0.75. For science, the average percentage of agreement was 83% with a corresponding kappa of 0.76. The criterion-related reliabilities were slightly higher: the average percentage of agreement for mathematics was 84% with a kappa of 0.78; for science, it was 86% with a corresponding kappa of 0.81. All of these values are above the accepted minimum thresholds established in the literature. The minimum threshold for the percentage of agreement is 80% and for Cohen’s kappa is 0.70. A table of all paired observations can be found in Appendix E.
Table 15

*Inter-Rater Reliability for Mathematics and Science*

<table>
<thead>
<tr>
<th></th>
<th>Percentage of agreement</th>
<th>Cohen’s kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics for 16 Paired Observations</td>
<td>81%</td>
<td>0.75</td>
</tr>
<tr>
<td>Science for 20 Paired Observations</td>
<td>83%</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 16

*Criterion-Related Reliability for Mathematics and Science*

<table>
<thead>
<tr>
<th></th>
<th>Percentage of agreement</th>
<th>Cohen’s kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics for 13 Paired Observations</td>
<td>84%</td>
<td>0.78</td>
</tr>
<tr>
<td>Science for 10 Paired Observations</td>
<td>86%</td>
<td>0.81</td>
</tr>
</tbody>
</table>

As part of the reliability measure, I examined the percentage agreement among paired observations for each construct. Considering reliability measures for single constructs could indicate possible improvement in the constructs or in the descriptive indicators, thereby yielding greater reliability overall. It could also indicate the need for additional information in specific areas during the training. Examining reliability for individual components of observation instruments is supported by other research studies (Applin, 2005; Marshall, Smart, & Horton, 2009; Sawada et al., 2002; Wright & Craig, 2011) and by the Standards for Educational and Psychological Testing (1999). The reliability measures for the individual constructs are listed in Table 17 and will be discussed in Chapter 5. The minimum threshold for acceptable reliability remains the same in this context and is 80% for percentage of agreement. Notable values include
three constructs in both mathematics and science that fell below the 80% threshold and the low percentage of agreement for the objective construct of 45% for science.

Table 17

**Percentage Agreement by Construct for Mathematics and Science**

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Agreements for Mathematics/Total for Math</th>
<th>Percent Agreement Mathematics</th>
<th>Agreements for Science/Total for Science</th>
<th>Percent Agreement Science</th>
<th>Percent Agreement for Entire Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>15/16</td>
<td>94%</td>
<td>9/20</td>
<td>45%</td>
<td>67%</td>
</tr>
<tr>
<td>Question Quality</td>
<td>15/16</td>
<td>94%</td>
<td>15/20</td>
<td>75%</td>
<td>83%</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>12/16</td>
<td>75%</td>
<td>20/20</td>
<td>100%</td>
<td>89%</td>
</tr>
<tr>
<td>Self-Evaluation</td>
<td>14/16</td>
<td>88%</td>
<td>20/20</td>
<td>100%</td>
<td>94%</td>
</tr>
<tr>
<td>Student Affect</td>
<td>10/16</td>
<td>63%</td>
<td>18/20</td>
<td>90%</td>
<td>78%</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>14/16</td>
<td>88%</td>
<td>13/20</td>
<td>65%</td>
<td>75%</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>11/16</td>
<td>69%</td>
<td>20/20</td>
<td>100%</td>
<td>86%</td>
</tr>
</tbody>
</table>

The Fleiss kappa was calculated as an inter-rater reliability measure when three or more observers rated the same class period. Due to logistic issues, no Facetime® observations using iPads® were conducted with more than two observers. However, two video recorded class periods—one for mathematics and one for science—were viewed by trained observers and the Fleiss kappa was calculated. The first video, Ms. Violet, was viewed and rated by three trained mathematics observers after the participants had conducted classroom observations using the AssessToday protocol. I had video recorded a mathematics lesson taught by Ms. Violet and
provided it to the observers for scoring. Since this lesson was recorded on a DVD, each observer could view the video on their own time and from different locations. The individual results were sent to me and I recorded and calculated the Fleiss kappa coefficient which was $k = 0.75$. This is considered substantial agreement (Landis & Koch, 1977).

The second video, Ms. Smith, was viewed and rated by 6 trained science observers after the participants had conducted several classroom observations using the AssessToday protocol. I used the *US2 Polymer* video from the Third International Mathematics and Science Study (TIMSS) for this observation (TIMSS, 1999). Five of the six participants viewed the video at the same time; however, they were instructed to rate the video individually without influence from others. The sixth observer viewed the video at a separate location and sent the results to me. The Fleiss kappa for the Ms. Smith video was $k = 0.73$, which falls into the substantial agreement range (Landis & Koch, 1977).

**Research Question 3**

Does exploratory factor analysis (EFA) identify any underlying factor structures in AssessToday?

I conducted an exploratory factor analysis (EFA) on the collected data, $n = 98$, to identify any underlying factor structures in AssessToday. Since AssessToday has seven constructs, 98 data points sufficiently satisfied the recommended number of participants per constructs—at least 5 and less than 20—by Henson and Roberts (2006), who additionally recommended conducting an EFA during instrument development in order to provide insights into the relationships between specific items. The loadings for the matrices were found using the computer program SPSS 20.
According to Keiffer (1999), one of the most widely used techniques for factor extraction in EFA is the principle component analysis method (PCA). This method was therefore used as an initial indicator of the basic structure. As was mentioned in chapter 3, the first step was to calculate the correlation matrix from the data. AssessToday contains seven constructs and thus a seven by seven correlation matrix was calculated (Table 18). The value of one on the long diagonal indicates that each construct is perfectly correlated with itself.

Table 18

**PCA Correlation Matrix**

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Objective</th>
<th>Question Quality</th>
<th>Nature of Question</th>
<th>Self Evaluation</th>
<th>Student Affect</th>
<th>Instructional Adjustment</th>
<th>Evidence of Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>1</td>
<td>.191</td>
<td>.096</td>
<td>.150</td>
<td>.042</td>
<td>.250</td>
<td>.157</td>
</tr>
<tr>
<td>Question Quality</td>
<td>.191</td>
<td>1</td>
<td>.621</td>
<td>.046</td>
<td>.433</td>
<td>.417</td>
<td>.263</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>.096</td>
<td>.621</td>
<td>1</td>
<td>.287</td>
<td>.536</td>
<td>.492</td>
<td>.340</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>.150</td>
<td>.046</td>
<td>.287</td>
<td>1</td>
<td>.236</td>
<td>.306</td>
<td>.286</td>
</tr>
<tr>
<td>Student Affect</td>
<td>.042</td>
<td>.433</td>
<td>.536</td>
<td>.236</td>
<td>1</td>
<td>.588</td>
<td>.189</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>.250</td>
<td>.417</td>
<td>.492</td>
<td>.306</td>
<td>.588</td>
<td>1</td>
<td>.349</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>.157</td>
<td>.263</td>
<td>.340</td>
<td>.286</td>
<td>.189</td>
<td>.349</td>
<td>1</td>
</tr>
</tbody>
</table>

The factor pattern matrix was then constructed from the correlation matrix (Table 19). The eigenvalues and the amount of variance were also calculated for each of the factors. The
eigenvalue is the amount of variance contained in each of the factors and is found by taking the sum of the squares of the loadings in each column.

Table 19

*Factor Pattern Matrix*

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Factor 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>.308</td>
<td>.583</td>
<td>.703</td>
<td>.217</td>
<td>.030</td>
<td>.137</td>
<td>-.063</td>
</tr>
<tr>
<td>Question Quality</td>
<td>.711</td>
<td>-.351</td>
<td>.352</td>
<td>-.202</td>
<td>.287</td>
<td>-.106</td>
<td>.336</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>.812</td>
<td>-.237</td>
<td>-.042</td>
<td>-.111</td>
<td>.322</td>
<td>.016</td>
<td>-.409</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>.456</td>
<td>.566</td>
<td>-.523</td>
<td>.234</td>
<td>.351</td>
<td>-.038</td>
<td>.139</td>
</tr>
<tr>
<td>Student Affect</td>
<td>.739</td>
<td>-.316</td>
<td>-.162</td>
<td>.335</td>
<td>-.238</td>
<td>.389</td>
<td>.093</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>.793</td>
<td>.036</td>
<td>-.002</td>
<td>.228</td>
<td>-.400</td>
<td>-.396</td>
<td>-.043</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>.546</td>
<td>.384</td>
<td>-.135</td>
<td>-.674</td>
<td>-.266</td>
<td>.119</td>
<td>.026</td>
</tr>
</tbody>
</table>

| Eigenvalue | 2.94 | 1.09 | 0.938 | 0.773 | 0.594 | 0.354 | 0.314 |
| % of variance | 42% | 15.5% | 13.4% | 11.0% | 8.5% | 5.1% | 4.5% |
| Total Variance | 42% | 57.5% | 70.9% | 81.9% | 89.4% | 95.5% | 100% |

Stevens (2002) recommended three methods to determine how many factors to retain in the interpretation. The first was to consider all factors with eigenvalues greater than one. Two factors met that criterion; factor one with eigenvalue of 2.94 and factor two with eigenvalue of 1.09.

The second method used for factor retention was the scree test. According to Stevens (2002), researchers should “retain all eigenvalues in the sharp descent before the first one on the line where they start to level off” (p. 389). It is clear from figure 11 that there is a distinct
change in the graph after factor 2, where the factors begin to level off. According to the suggested criterion, the scree plot visually supports two factors (Figure 11).

Figure 11. The scree plot for the PCF indicates two strong eigenvalues (factors 1 and 2) and a distinct point where the eigenvalues begin to level off.

A third method for factor retention is the 70% rule. Stevens (2002) suggests retaining as many factors as will account for at least 70% of the variance. The total variance accounted for can be found in Table 19 and would indicate retaining at least three factors instead of the two implied by the eigenvalue greater than one method and the scree test. Retaining three factors would account for 70.9% of the variance.

To interpret the meaning of the factors and determine what latent construct might underlie them, I first examined the magnitude of the loadings in the factor pattern matrix (Table 19) for each retained factor. Keiffer (1999) suggested considering all items (constructs) that load with a coefficient greater than |0.30|. For factor one, all seven constructs met the criterion and loaded with values that ranged from 0.308 to 0.812. With all of the constructs loading together, the first factor seemed to indicate that all the constructs were addressing one central concept. Since the constructs were established from the formative assessment literature and
confirmed by the foremost experts in the field, the one central concept that loaded on factor one can be considered as formative assessment, which is the overarching theme of AssessToday.

Factor two consisted of both positive and negative loadings (Table 19). Stevens (2002) describes this as a “bipolar factor” (p. 387). Stevens suggests that a contrast between two latent constructs was indicated. For the second factor, three constructs loaded that were positive and greater than 0.30; objectives, self-evaluation, and evidence of learning. Two constructs loaded that were negative and less than -0.30; question quality and student affect (Table 19). Nature of questioning was also negative and fell just below the |0.30| threshold recommended by Keiffer (1999).

The formative assessment literature suggests that formative assessment consists both of teacher actions and student actions. This secondary factor can be explained as a contrast between the two actions. For students, the objectives indicate that they know the learning targets for the class period and are striving to attain them; self-evaluation suggests they are engaged in the learning process, and evidence of learning reveals whether the student has successfully demonstrated mastery of the stated learning target. In contrast, the other two constructs focus on the teacher’s actions. Teachers should be proficient questioners and should be cognizant of the affective needs in the classroom.

Factor 3 is another bipolar factor where self-evaluation (loading of -0.523) is contrasted with the objectives (loading of 0.703) and question quality (loading of 0.352) constructs. I am unsure what underlying latent construct factor three is indicating, especially since objectives and self-evaluation loaded together in factor 2 and yet were contrasted in factor 3. According to Keiffer (1999) different retention methods often generate different results. It is the responsibility of the researcher to determine how many factors to retain based on the theoretical framework. I
therefore believe that AssessToday is best described by only two factors. To help develop a clear interpretation, I investigated the rotated factor pattern matrix as recommended by Keiffer (1999) and Stevens (2002).

Keiffer (1999) stated that “interpretation of the factor results is almost always aided by the rotation of the factor solution” (p. 80). Although there are several methods available for factor rotation, orthogonal varimax rotation is one of the most popular and its use is suggested by Keiffer and Stevens (2002). This method was conducted in order to redistribute common variance and consider the strength of each construct as it relates directly to the factors. Additionally, Stevens (2002) suggests that this method will “clean up the factors” (p. 391) and make it easier to interpret them. Varimax rotation produces factors that have strong loadings for a small number of constructs and near zero loadings for the others. The orthogonally rotated data (Table 20) provided strong support for my interpretation of factor 1.

Table 20

Rotated Principle Component Matrix with Varimax Rotation

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Factor 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>.078</td>
<td>-.004</td>
<td>.063</td>
<td>.065</td>
<td>.988</td>
<td>.097</td>
<td>.017</td>
</tr>
<tr>
<td>Question Quality</td>
<td>.922</td>
<td>.179</td>
<td>.111</td>
<td>-.023</td>
<td>.096</td>
<td>.153</td>
<td>.270</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>.324</td>
<td>.241</td>
<td>.154</td>
<td>.142</td>
<td>.020</td>
<td>.188</td>
<td>.870</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>-.015</td>
<td>.089</td>
<td>.129</td>
<td>.973</td>
<td>.068</td>
<td>.111</td>
<td>.104</td>
</tr>
<tr>
<td>Student Affect</td>
<td>.183</td>
<td>.916</td>
<td>.054</td>
<td>.102</td>
<td>-.008</td>
<td>.264</td>
<td>.210</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
<td>.168</td>
<td>.286</td>
<td>.160</td>
<td>.139</td>
<td>.130</td>
<td>.893</td>
<td>.177</td>
</tr>
</tbody>
</table>
The orthogonal rotation of the factor structure matrix provided additional support for a single overarching construct for AssessToday that fits with the theoretical framework and the first factor of the factor structure matrix. All the constructs of AssessToday are supported by strong loadings on individual factors (Table 20). Taken together, the overarching theme of formative assessment is clearly seen by combining the individual factors into a single unified interpretation.

Since the orthogonal rotation of the factor structure matrix did not provide supporting evidence for the interpretation of factor 2, I conducted further investigations to see if the bi-polar structure was supported by other factor extraction techniques. Henson and Roberts (2006) suggested EFA should consider as many components of factor analysis as possible in order to collect a complete picture of the factors with their interconnections and structure. The principle axis factoring method with varimax rotation provided additional support (Table 21). This method is supported by Keiffer (1999) and Stevens (2006) and is considered as one of the two most common methods used by researchers.
Table 21

*Principle Axis Factor (PAF) Matrix with Varimax Rotation*

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectives</td>
<td>.068</td>
<td>.044</td>
<td>.089</td>
<td>.554</td>
<td>.085</td>
<td>.002</td>
</tr>
<tr>
<td>Question Quality</td>
<td>.765</td>
<td>.251</td>
<td>-.120</td>
<td>.224</td>
<td>.169</td>
<td>-.103</td>
</tr>
<tr>
<td>Nature of questioning</td>
<td>.738</td>
<td>.315</td>
<td>.292</td>
<td>-.017</td>
<td>.185</td>
<td>.156</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>.029</td>
<td>.158</td>
<td>.618</td>
<td>.124</td>
<td>.189</td>
<td>.011</td>
</tr>
<tr>
<td>Student affect</td>
<td>.351</td>
<td>.743</td>
<td>.183</td>
<td>-.055</td>
<td>.010</td>
<td>-.088</td>
</tr>
<tr>
<td>Instructional adjustment</td>
<td>.211</td>
<td>.687</td>
<td>.155</td>
<td>.297</td>
<td>.301</td>
<td>.148</td>
</tr>
<tr>
<td>Evidence of learning</td>
<td>.192</td>
<td>.108</td>
<td>.240</td>
<td>.135</td>
<td>.516</td>
<td>.006</td>
</tr>
</tbody>
</table>

The bi-polar structure suggested in the discussion of the PCA can be clearly seen in table 21 by combining factors 1 and 2 as teacher actions and factors 3, 4, and 5, as student actions. Other notable observations include:

- The loadings of 0.765 on *Question Quality* and 0.738 on *Nature of Questioning* fall together into one factor; questioning. This is consistent with the literature which suggests questioning is a very important component in the classroom. Teachers who
are consistent questioners will provide good question quality and structure. Similarly, teachers who are ineffective as questioners will be weak in both areas. The questioning literature and practical experience suggests these should remain separate constructs for the instrument to assess. The rationale is that for teachers to be effective questioners they need both quality questions and good question structure (Black & Harrison, 2001; Wilen & Clegg, 1986). The AssessToday will delineate between the two and be able to offer teachers specific guidelines as to how to improve their questioning skills.

- The loadings of 0.743 for student affect and 0.687 for instructional adjustments fall together into one factor; feedback. This is consistent with the formative assessment literature and practical experience, since effective users of formative assessments base their instructional adjustments upon the feedback they receive from their students.

- Self-evaluation, Objectives, and Evidence of Learning stand alone as unique sub-factors supporting the overall factor structure of formative assessment.

It is the conclusion of the researcher that the EFA conducted on the data collected from AssessToday observations reveals two underling structures. The first is simply a confirmation that AssessToday does measure the single overarching construct of formative assessment. The supporting evidence for this conclusion comes from the first factor of the PCA and the rotated factor pattern matrix. For the first factor of the PCA, every construct of AssessToday loaded with notable strength greater than the |0.30| threshold. This structure was confirmed through the rotated factor pattern matrix since each construct loaded strongly on individual factors indicating strong correlations between the individual constructs and single factors as seen in table 20.
AssessToday was designed to score a teacher’s use of short-cycle formative assessment and the results from the factor analysis support that proposition.

The second underlying structure revealed from the EFA is the bipolar structure indicated in the second factor of the PCA (Table 19) and supported by the PAF (Table 21), which is the contrast between a teacher’s actions and the students’ actions. The literature review also supports the view that formative assessment consists of actions from both teachers and students.

Research Question 4

Does AssessToday display evidence of construct validity as determined by content validity, the EFA, and reliability?

Establishing construct validity requires three steps: (a) a set of theoretical concepts and their interrelations, (b) a method to measure and examine the theoretical constructs developed, (c) and empirical evidence to support the proposed constructs (Carmines & Zeller, 1979, Chronbach & Meehl, 1955). For AssessToday, its concepts and their interrelations were developed through an examination of formative assessment literature. The constructs are well supported and widely accepted (Black et al., 2003; Black & Wiliam, 2009; Brookhart, Moss & Long, 2008; Heritage, 2010; McManus, 2008; Popham, 2008; Stiggins, 2009; Wiliam et al., 2004). The development of how to measure the constructs was guided by the Standards for Educational and Psychological Testing (1999), empirical research studies on similar instruments (Marshall, Smart, & Horton, 2009; O’Malley et al., 2003; Sawada et al., 2002; Wright & Craig, 2011), and experts in the field of assessment (Bresciani et al., 2009; Cantor, 1996; Cohen, 1960; Fleiss, 1971; Nichols et al., 2011; Smith, Vannest & Davis, 2011).

AssessToday was empirically tested in over 90 classrooms, on 56 teachers, with 12 observers. Inter-rater reliabilities were sufficiently robust for satisfactory reliability measures.
Although three of the constructs in mathematics and three of the constructs in science fell below the associated threshold of 80%, the instrument as a whole produced moderate to substantial agreement. The EFA produced clear evidence to support validity.

Construct validity is defined by Haynes (2001) as: “The evidence and rationales indicating the degree to which data from an assessment instrument measures the targeted construct” (p. 75). First of all, university professors, practitioners, and experts in the field of formative assessment agree that the constructs contained in AssessToday are both appropriate and comprehensive for gathering data on short-cycle formative assessment. Second, the data collected from the mathematics and science observations serve as evidence for construct validity and are discussed in more detail in Chapter 5. Finally, the principle component analysis, the scree plot, and the principle axis factoring with varimax normalization rotation provide sufficient evidence that AssessToday measures what it claims to measure (formative assessment) and provides additional insights into the interrelationships of the constructs.
CHAPTER 5
DISCUSSION AND CONCLUSIONS

This chapter presents a discussion of the results of a validity and reliability study for mathematics and science conducted on the AssessToday observation instrument. AssessToday is an observation protocol that was designed to measure a teacher’s use of short-cycle formative assessment in a single class period. The constructs and the associated indicators for AssessToday were suggested by an examination of the formative assessment literature as described in Chapter 2. The scoring scale and design of the instrument are supported by the Standards for Educational and Psychological Testing (1999), as well as several empirical studies on validation and reliability measures for observation instruments (Appledorn, 2004; Applin, 2005; Clifford et al. 2009; O’Malley et al. 2003; Marshall, Smart, & Horton, 2009; Sawada et al., 2002; Wright & Craig, 2011).

The purpose of the study was four fold. First, the study sought to determine if AssessToday demonstrated content validity as determined by experts in the field of formative assessment and professionals in the discipline of education. Second, this study analyzed inter-rater reliability coefficients calculated from paired observations of trained observers to determine if AssessToday produced reliable results. Third, an EFA was conducted to determine the factor structure of AssessToday. Finally, the study examined evidence of construct validity as determined by content validity, the EFA, and reliability.

The current study on AssessToday fills an important gap in the development of short-cycle formative assessment. The discussion of the results of this study will be presented in the same order the research questions were posed. First, the I will present a discussion on the evidence supporting the content validity of AssessToday. Second, the reliability coefficients
calculated from the paired observations are addressed. Third, I draw conclusions from the EFA. Finally, I discuss evidence for the construct validity of the AssessToday protocol. The chapter concludes with the limitations of the current study and suggestions for future research related to AssessToday and its use in educational settings.

Discussion of Research Question 1

The first research examined if AssessToday demonstrate content validity as determined by experts in (a) the field of formative assessment and (b) professionals in the field of science and mathematics education?

I believe that the results from the three-part process presented in Chapter 4 support the content validity of AssessToday. The first process discussed was a thorough review of the formative assessment literature, the second was a comparison of the most important elements of short-cycle formative assessment with the constructs of AssessToday, and the third was to consider feedback from experts in the field of formative assessment and professionals in mathematics and science education.

This three part process supports content validity in two ways. First of all the literature review provides evidence that AssessToday is connected to and developed from a body of research that describes the theoretical framework of short-cycle formative assessment. The literature review covers a 23 year period, from Sadler (1989) until the present day and describes the growth and development of formative assessment over those years. Sadler provided the basic framework with three questions students should ask: *Where am I now?*, *Where am I going?*, and *How do I get there?* Other researchers provided additional details that added to the basic structure that was established by Sadler. AssessToday was gleaned from the work and insights of those researchers (See Table 5).
The feedback from experts, practitioners, and professionals provides a second source of evidence for content validity. Although they suggested modifying the construct indicators, they agreed that AssessToday included everything it should, and nothing it should not. This is one of the conditions for content validity recommended by AERA, APA, and NCM (1999). The most significant feedback was obtained from the five experts; Brookhart, Heritage, McMillan, Popham, and Stiggins. These five researchers are leaders in the field of formative assessment and to conducting professional development with teachers and administrators. According to these foremost experts, the constructs of AssessToday sufficiently encompass the theoretical framework of short-cycle formative assessment. Their input as professionals provides strong evidence for the content validity of the AssessToday observation instrument.

In addition to the three part process mentioned above, trained observers field tested AssessToday in order to gain information about the instruments appropriateness for classroom use. Although some modifications to the indicators were recommended, the observers were sufficiently satisfied with the formative assessment constructs and the ease with which to score the instrument. The number of constructs was manageable and distinguishable during a single class period. According to the trained observers, AssessToday was useful and appropriate for use in the classroom. The practical use of the instrument provides additional support for the content validity of AssessToday as suggested by Haynes (2001).

Finally, I considered the distribution of the teachers’ scores among the four levels of proficiency for each of the constructs (figures 2 – 10). This information was significant for two reasons. First of all, the distribution indicated an overall weakness in formative assessment use among the participants in this study. This can be useful information for both teachers and administrators. Teachers are able to ascertain needed areas of improvement in their own
teaching and administrators are able to notice areas for professional development. Professional learning communities would also benefit from this type of information. Secondly, the distribution can inform teacher educators on possible weaknesses in university instruction. I am also a teacher educator and am interested in improving the quality of teachers entering the teaching profession. Information of this type that was gathered from teachers who were previous students could be a source of information that could improve teacher education.

Discussion of Research Question 2

The second research question examined if AssessToday demonstrated reliability in science and mathematics as determined from inter-rater reliability coefficients of trained observers of AssessToday?

Inter-rater agreement and consistency for trained observers are important in the consideration of reliability coefficients (Brandon et al., 2008). Smolkowski and Gunn (2010) required their observers to maintain an 80% rate of agreement or higher before they were allowed to observe teachers as did Downer et al. (2010). The reliability study of AssessToday followed similar guidelines and was established in two ways. First, observers were trained to use the instrument in an appropriate manner and the consistency of their scores were compared with established norms. Practical use in the classroom has shown that if observers follow the established protocols as described in the handbook, observer pairs can agree on at least five out of the seven constructs on a consistent basis. Chatterji (2003) suggests that even with ample training and a thoroughly designed rubric for scoring, human error will occur. She recommends maintaining several observers that researchers can draw from in order to select those who are more consistent than others. In the case of the current research, two trained observers were not able to demonstrate the level of consistency required and their corresponding data from paired
observations were not included in the study. These observers were excluded in an effort to maintain the integrity of the study at the highest level possible (Wolfe, Koa & Ranney, 1998).

Chatterji (2003) posits one possible explanation for inconstant scores is rater bias and Wolfe, Koa, and Ranney (1998) suggest the raters probably have a different understanding of the scoring criteria. Researchers argue that additional training with the inconsistent scorers could help generate more valid and reliable scores (Chatterji, 2003; Wolfe, Koa, & Ranney, 1998). Due to the time constraints of this study, additional training was not provided for the two observers who seemed to have difficulty with reliability in terms of applying the scoring criteria to an observation.

Second, I provided three measures of inter-rater reliability as reliability evidence: percentage of agreement between raters, Cohen’s kappa, and Fleiss kappa. The reliability coefficients for AssessToday are sufficiently robust to support the reliability of AssessToday. For paired observations, Carmines and Zellar (1979) suggest a minimum of 80% agreement for the percentage of agreement measure and a minimum of $k = 0.7$ for Cohen’s kappa. As stated in Chapter 4, the average percentage of agreement for mathematics was 81% with a corresponding kappa of $k = 0.75$ and the average science percentage of agreement 83% with a corresponding kappa of $k = 0.76$ (Table 15). The criterion-related reliability was slightly higher with the percentage of agreement for mathematics at 84%, with kappa of $k = 0.78$, and the percentage of agreement for science at 86%, with kappa of $k = 0.81$ (Table 16). AssessToday produced coefficients that were above recommended thresholds; however, they were at the lower end. For observations with three or more observers, Landis and Koch (1977) recommend a minimum threshold of $k = 0.7$. When using the Fleiss Kappa, a coefficient of $k = 0.75$ was determined for mathematics and $k = 0.73$ for science. Similar to the paired observations, both Fleiss kappa
coefficients are at the lower end of the recommended threshold. Evidence for reliability is provided through these values, but at the same time care should be taken to continue improvements and revisit reliability measures in the future.

Although the reliability coefficients for the entire instrument were sufficiently robust, the analysis did reveal some weaknesses in specific construct reliability (Table 17). For mathematics, the paired observations conducted indicated three of the constructs of AssessToday produced reliability measures that were below the 80% threshold. The following percentages were produced: nature of questioning 75%, student affect 63%, and evidence of learning 69%. The other four constructs scored well above the 80% threshold ranging from 88% to 94%.

Similarly in science, three of the constructs yielded scores below the approved level. Objectives scored 45%, question quality yielded 75%, and instructional adjustments scored 65%, with the other four constructs ranging from 90% to 100% agreement. When the percentage of agreement for mathematics and science were combined for each construct, the scores ranged from 67% to 94% with three constructs scoring below 80%. Objectives scored 67%, instructional adjustments 75%, and student affect 78%. These scores indicate areas of focus for the further development of AssessToday.

When considered together, all the constructs of AssessToday met the recommended reliability measure at some point in the study. Three of the constructs in mathematics and three of the constructs in science failed to meet the recommended measure; however none of the constructs fell below the recommended measure of 80% for percentage of agreement and k = 0.7 for the kappa statistics in both mathematics and science. This does indicate that all constructs of AssessToday are worded sufficiently and specific enough for observers to provide consistent and
reliable results. At the same time, these values also indicate that there is still room for improvement and suggestions are provided later in this section.

Jonsson and Svingby (2007) suggest considering holistic scores as well as exact agreement in order to provide a measure of rater consistency. This less stringent measure of reliability is calculated by considering how many scores differ by only one point or less in paired observations. Downer et al. (2010) followed the same protocol and required their observers to score within one point of each other on 80% of their scores as did Wright and Craig (2011). For the current study, zero scores in mathematics were more than one point away on paired observations and only two scores in science were more than one point away. Both science scores that were two points away from each other were from the objective construct and were given by the same observer. I also noticed that this same observer differed on the objective construct on all of their paired observation scores. This indicates that either the observer misunderstood the indicators for the construct or lost focus during the observations and did not observe the associated action. This is significant to notice since the science objective construct received the lowest level of inter-rater agreement at 45%. If the percentage of agreement for the objective construct was calculated without the scores from this one observer, the reliability would increase to 60%. Although the removal of scores for this one observer would not increase the objective reliability to 80%, it does emphasize the importance of rater training and focus.

Improving the Reliability of AssessToday

To increase the reliability of AssessToday, all the constructs should be examined further to increase clarity and decrease subjectivity. Although all the conditions for reliability and validity have been met, additional improvements can be achieved through continual consideration of the constructs and supporting indicators. The trained observers who
participated in this study did provide written feedback during their observations. Their insight as trained users of the instrument can shed light on specific words or phrases that are problematic.

One example was a comment made by a trained mathematics observer who had a question about the objective construct. The observer wrote, “What does it look like when a teacher revisits the objectives?” Upon examination of the training material, no specific information was presented to address this concern. A science rater asked, “What is the difference between journaling and note taking?” This question might indicate the observer was not familiar with the term journaling. Another observer was confused about the difference between the question quality and the nature of questioning constructs. In the explanation of their score, the observer used the indicators of question quality to justify their score of the nature of questioning. These are all indicators that some confusion about the constructs was present among the trained observers. Although video examples were presented during the training, more examples and discussion would be beneficial for a clearer understanding of the scoring process. A list of commonly asked questions with appropriate answers would also be beneficial during the observer training.

Another way to improve the reliability of AssessToday is to expand the training of the observers. Although the amount of preparation yielded valid and reliable results, additional instruction will improve the reliability coefficients and observer confidence. Other researchers dedicated from 2 hours to 3 days to train their observers (Applin, 2005; Brandon et al., 2008; Downer et al., 2010; O’Malley et al., 2003; Smolkowski & Gunn, 2012). One researcher (O’Malley et al., 2003) used “brainstorming” during training to allow the participants an opportunity to think of ways the constructs could be seen in the classroom (p. 92). Downer et al., (2010) required their trainees to observe 5 video clips and compare scores with an instructor. In
addition to additional video clips of specific constructs, paired observations should also be conducted in the classroom until the novice is able to reach 80% agreement with an expert (Smolkowski & Gunn, 2012). In the current study, only one video was viewed and rated during the training process and no conditions pertaining to additional observations were placed on the trainee. As noted earlier, observation scores from two observers were removed from the study due to a lack of consistency. Had conditions similar to those mentioned above been in place during the training, appropriate guidance could have been provided to the trainees to improve their understanding of the constructs.

Discussion of Research Question 3

The third research question explored if exploratory factor analysis (EFA) identified any underlying factor structures in AssessToday?

According to Henson and Roberts (2006), “[f]actor analysis can be used to determine what theoretical constructs underlie a given data set and the extent to which these constructs represent the original variables” (p. 396). Results from the EFA support the theoretical constructs of formative assessment and provide evidence of how those constructs are interwoven (Figure 11). The primary factor from the principle component analysis (PCA) matrix supports a single overarching interpretation thereby unifying the seven constructs of the instrument and is thus interpreted as formative assessment. This conclusion is supported by the rotated factor pattern matrix. In addition, two sub-factors are clearly delineated and are interpreted as teacher action and student action. The principle axis factoring (PAF) matrix provided additional clarity to the interpretation of the second factor of PCA. Student actions consists of three constructs - objectives, self-evaluation and evidence of learning. Teacher actions consists of two sub-levels - questioning and feedback, with questioning being composed of question quality and nature of
questioning and feedback being composed of student affect and instructional adjustments.

Figure 12 represents the nested structure of AssessToday as provided by the conducted factor analysis. The internal structure described through the factor analysis agrees with the constructs developed through the theoretical framework and thus provides evidence for the construct validity of AssessToday.

<table>
<thead>
<tr>
<th>Formative Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Responsibility</td>
</tr>
<tr>
<td>Student Responsibility</td>
</tr>
<tr>
<td>Questioning</td>
</tr>
<tr>
<td>Quality</td>
</tr>
<tr>
<td>Structure</td>
</tr>
<tr>
<td>Feedback</td>
</tr>
<tr>
<td>Student Affect</td>
</tr>
<tr>
<td>Instructional Adjustments</td>
</tr>
<tr>
<td>Objectives</td>
</tr>
<tr>
<td>Self – Evaluation</td>
</tr>
<tr>
<td>Evidence of Learning</td>
</tr>
</tbody>
</table>

*Figure 12. AssessToday factor structure from the EFA. Factor 1 was interpreted as formative assessment, factor two was bipolar and indicated a contrast between teacher and student responsibility.*

**Discussion of Research Question 4**

The fourth research question examined if AssessToday displayed evidence of construct validity as determined by content validity, the EFA, and reliability?

In the previous discussion of Questions 1, 2, and 3, I examined evidence that supported content validity and reliability as well as analyzed the EFA that was conducted on the data. According to Messick (1995), construct validity is “an integration of any evidence that bears on
the interpretation of test scores including content validity and criterion-related evidence” (p.
742). In addition, Haynes (2001) describes construct validity as, “the evidence and rationales
indicating the degree to which data from an assessment instrument measures the targeted
construct” (p. 75). The data from this study does support the construct validity of AssessToday
and is discussed in the following section.

First of all, the strong evidence from content validity contributes to the construct validity
of AssessToday. The instrument is firmly rooted in formative assessment literature and experts
from several content areas have examined and approved the constructs and associated indicators.
Second, the inter-rater reliability coefficients also provide evidence of construct validity.
Although some of the individual constructs fell below the 80% threshold, the overall inter-rater
reliability coefficients for both mathematics and science were well within acceptable standards.
Thirdly, the criterion-related reliabilities were also sufficient to provide evidence for construct
validity. Finally, the results from the EFA suggest that AssessToday is a unified measure of
short-cycle formative assessment, and the constructs measure its components appropriately.

Litwin (1995) posits that construct validity is also a measure of how meaningful and
useful an instrument is when in practical use. AssessToday is very meaningful and useful in the
classroom context. In the first place, AssessToday is a protocol designed to inform teachers
about their use of formative assessment in the classroom. Research has demonstrated that if
teachers improve their intentional use of formative assessment, then it is expected that the
achievement of their students will also improve (Black, 2004; Black & Wiliam, 1998; Meisels et
al., 2002; Rodriguez, 2004). Second, the instrument was designed with practitioners in mind.
The words and phrases found in the protocol were intentionally chosen to communicate with
educators. Third, the current study demonstrated that reliable and valid results can be obtained
from approximately two hours of training. Thus, AssessToday is a meaningful and useful
instrument and therefore provides supporting evidence for construct validity.

AssessToday is a trustworthy instrument that will return valid and reliable scores with
proper use and training. Researchers should continue to search for additional evidence that will
provide greater support for construct validity in the field of education.

Improving the construct validity of AssessToday

To improve the validity of AssessToday, additional validity studies should be conducted.
Haynes (2001) states that “construct validity is the sum of [the] inferences from multiple validity
indices” (p. 77). He then lists 13 different indices that comprise construct validity researchers
should address. The current study considered three of these indices; internal consistency, content
validity, and inter-observer reliability. These were chosen in keeping with the basic validity
be ascertained after the instrument has been applied in multiple settings with diverse populations
over several years. The current study considered the application of AssessToday in middle
school and high school mathematics and science classes over a period of approximately one year.
The current study was also conducted with participants involved with a professional
development program at a large university in Texas. Much is left to do; however, the current
study began the process of discovering evidence to support the validity of AssessToday. Future
studies should address not only the other indices listed by Haynes but also other content areas,
age groups, and localities.

Noteworthy Observations Made During the Study

Although the current study was designed to examine the validity and reliability of
AssessToday, other interesting outcomes resulted from the research process. First of all, I
noticed on a few occasions during the observations, that teachers either taught content incorrectly or did not use appropriate scaffolding for learning. On one occasion, for example, a teacher informed a student that a person could not divide by a fraction but rather needed to change the fraction to a decimal and use long division. Although the method presented to the student was correct, the information about dividing by a fraction was incorrect. The teacher missed an opportunity to scaffold deep content knowledge, build unity and consistency in mathematical understanding, and discouraged further inquiry from her students. An observer who was not a content specialist might have missed this mathematical misunderstanding, and consequently an opportunity to improve the instruction in that classroom.

Although AssessToday is not a content specific instrument, important elements in the classroom will be missed if the observer is not a content specialist. The high level for the question quality construct, for example, defines it as when a teacher, “consistently and appropriately uses questions effectively to scaffold instruction.” If the observer does not have a deep understanding of the content, then the ability to judge questioning as appropriate or effective in the scaffolding of instruction will be diminished. A second construct which a non-expert observer could have difficulty with is instructional adjustment. Without the appropriate background in content, the observer will not be able to ascertain an appropriate adjustment in instruction.

A second but equally important observation I made was the interesting distribution of scores among the constructs for mathematics and science (See Figures 4 – 10) and possible conclusions that can be drawn from the data. For the individual constructs, the data clearly reveal deficiencies of formative assessment use for both mathematics and science teachers. For example, over half the scores for both mathematics and science teachers ranked in the lower two
levels for self-evaluation and instructional adjustments. Clearly, this cohort of teachers would benefit from professional development in these two areas. There are also noticeable similarities and differences between the two content areas. One difference is that science teachers perform significantly better with objectives than mathematics teachers (Figure 4). The study showed that 44% of science teachers scored moderately high to high level, whereas only 13% of mathematics teachers scored in that range. Mathematics teachers, on the other hand, are slightly better at questioning with 55% scoring in the top two levels and 52% of science teachers (Figures 5 and 6). The data from the other 4 constructs indicate mathematics and science teachers are very similar. Overall, science teachers scored higher on a composite score of all seven constructs with 48% of the scores in the top two levels compared with 39% of mathematics teachers.

A third observation was the surprising number of teachers who scored in the lowest level of formative assessment use in the composite score. The results showed 8% of mathematics teachers and 10% of science teachers were rated in the lowest level. This is somewhat alarming as this indicates there are some teachers who are completely deficient in certain aspects of formative assessment use in the classroom. On the other hand, this information can be used to inform teachers of their deficiencies and provide them with the opportunity for improvement.

The observations made above are just a few examples of the practical and useful insights AssessToday can provided for teachers and administrators. Teaching is a complex task and discovering areas for personal improvement can reap great benefits in the classroom.

Areas for Future Research

This study has only begun part of the research that is possible with the AssessToday observation instrument. Of first importance seems to be additional validity studies, especially concurrent validity with other observation instruments as recommended by Messick (1995) and
Haynes (2001). Validity studies in other content areas, age groups, and demographics are needed along with longitudinal studies of formative assessment use and development. These studies would not only support construct validity, they would also reinforce the generalizability of the instrument. Data should be collected from the general population of mathematics and science teachers to see if similar results would be obtained. Future research might also include a comparison between a teacher’s impression of how they performed and how an observer (3rd party) viewed their teaching. This would connect teacher efficacy and formative assessment use. Further, comparative studies between public and private institutions as well as among different countries, states, or districts might also provide useful insights. Finally, studies could be conducted on the relationship between a teacher’s use of feedback obtained from AssessToday and student achievement in the classroom.

Limitations of the Study

One of the limitations of this study is the content specific nature of the observations that were conducted. As noted in the standards (AERA, APA, & NCM, 1999), it is important for researchers to determine content validity and reliability for each subgroup in the larger population. The current study examined the validity and reliability of AssessToday in middle school and high school mathematics and science classrooms.

A second limitation of the current study is the age restrictions imposed. The AssessToday validation and reliability coefficients apply to middle school and high school only. Further research would be necessary in elementary mathematics and college mathematics to determine if AssessToday is both valid and reliable in those contexts.

A third limitation of the current study is the specific nature of the participants in the research. All participants are current or former teachers in the same mathematics and science
professional development. These participants received professional development training in mathematics or science pedagogy along with specific content enrichment. Due to this special training, it is possible the scores of the participants will be slightly higher than scores in the general population of educators.

A fourth limitation is the small number of observations conducted. Although over 100 observations were conducted during the course of this study, this is still a relatively small number in comparison to the number of classes taught each year. In addition, teachers’ scores were based on single observations. Observing teachers multiple times would provide a clearer picture of their use and understanding of formative assessment.

Conclusion

There is a general consensus that most schools need to improve the quality of education in the classroom. Teachers, administrators, and policy makers have tried various programs and incentives to accomplish this elusive goal. Although there may be several strategies that are able to make this goal a reality, formative assessment is a proven method that can increase achievement and motivation when used correctly. AssessToday was developed to inform teachers about their formative assessment use and thus provide a framework for improvement.

The current study examined the validity and reliability of AssessToday among middle school and high school mathematics and science teachers. The results of the study revealed strong content validity, robust inter-rater and criterion-related reliabilities, appropriate results from exploratory factor analysis, and therefore support for construct validity. Although there is room for improvement, AssessToday has been proven to be a valid and reliable measure of a teacher’s implementation of short-cycle formative assessment when used appropriately. The development and subsequent validation of AssessToday will provide an excellent tool to assist
teachers in their development of formative assessment in the classroom. The creators did not
design this instrument as an evaluative tool to be used as judgments on teachers; rather, it is an
informative tool to assist teachers in their personal development. It is my hope that AssessToday
will be used by teachers and professional learning communities for this purpose.
APPENDIX A

ASSESSTODAY OBSERVATION INSTRUMENT
<table>
<thead>
<tr>
<th>Core Constructs of Formative Assessment</th>
<th>Low</th>
<th>Moderately Low</th>
<th>Moderately High</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective(s) (Clear Targets)</td>
<td>Zero or one of the indicators is observed during the class period.</td>
<td>Two of the four indicators are observed during the class period.</td>
<td>Three of the four indicators are observed during the class period.</td>
<td>Teacher states, writes and restates objective(s) throughout the lesson. Students write down the objective(s).</td>
</tr>
<tr>
<td>Question Quality</td>
<td>Teacher does not use questions effectively during instruction.</td>
<td>Teacher rarely uses questions effectively to scaffold instruction.</td>
<td>Teacher generally uses questions effectively to scaffold instruction.</td>
<td>Teacher consistently and appropriately uses questions effectively to scaffold instruction.</td>
</tr>
<tr>
<td>Nature of Questioning</td>
<td>Teacher does not use questioning effectively during instruction.</td>
<td>Teacher rarely uses questioning effectively to diagnose problems with learning and improve instruction.</td>
<td>Teacher generally uses questioning effectively to diagnose problems with learning and improve instruction.</td>
<td>Teacher consistently and appropriately uses questioning effectively to diagnose problems with learning and improve instruction.</td>
</tr>
<tr>
<td>Self-evaluation</td>
<td>Teacher does not use student self-evaluation during instruction.</td>
<td>Generic self-evaluation is employed but the strategy is not explicitly tied to the regulation and improvement of student’s self-learning.</td>
<td>Evidence of one self-evaluation strategy is used during instruction in an effort to regulate and improve the student’s self-learning. Self-evaluation strategy could include techniques such as the use of traffic lights, a self-assessment inventory, journaling, or a reflection statement.</td>
<td>Teacher uses a variety (two or more) of strategies to encourage students to self-evaluate in an effort to regulate and improve their own learning. These could include techniques such as the use of traffic lights, a self-assessment inventory, journaling, and/or reflection statements.</td>
</tr>
<tr>
<td>Instructional Adjustment</td>
<td>No adjustments to instruction are observed.</td>
<td>Teacher uses minimal adjustments during instruction.</td>
<td>Teacher predominately and effectively uses adjustments during instruction.</td>
<td>Teacher consistently and effectively uses adjustments during instruction.</td>
</tr>
<tr>
<td>Evidence of Learning</td>
<td>There is minimal evidence that learning occurred. There is evidence of learning for 0-24% of the students.</td>
<td>Token use of individualized assessment strategy. There is evidence of learning for 25-49% of the students.</td>
<td>Moderate individual assessment strategy. At least one other group strategy is used (i.e. small group instruction, whole group instruction, or peer feedback) There is evidence of learning for 50-74% of the students.</td>
<td>Near total individualized assessment strategy. At least one other group strategy is used (i.e. small group instruction, whole group instruction, or peer feedback.) There is evidence of learning for 75-100% of the students.</td>
</tr>
</tbody>
</table>
APPENDIX B

CONSTRUCTS WITH INDICATORS
Objectives Construct with Indicators

High-level - Teacher states, writes, and restates the objective(s) throughout the lesson. Students write down the objective(s).

- Teacher states objective(s).
- Teacher writes objective(s) for students to see.
- Students write objective(s) for themselves (80-100% of students).
- Teacher revisits the objective(s) during the class period.

Moderately-high – Three of the four indicators are observed during the class period.

- Teacher states objective(s).
- Teacher writes objective(s) for students to see.
- Students write objective(s) for themselves (80-100% of students).
- Teacher revisits the objective(s) during the class period.

Moderately-low – Two of the four indicators are observed during the class period.

- Teacher states objective(s).
- Teacher writes objective(s) for students to see.
- Students write objective(s) for themselves (80-100% of students).
- Teacher revisits the objective(s) during the class period.

Low-Level – Zero or one of the indicators is observed during the class period.

- Teacher states objective(s).
- Teacher writes objective(s) for students to see.
- Students write objective(s) for themselves.
- Teacher did not revisit the objective(s) throughout the lesson.
Question Quality Construct with Indicators

High-level - Teacher consistently and appropriately uses questions effectively to scaffold instruction.

- Open ended questioning – Teacher consistently and appropriately uses open-ended questions that challenge students to think critically during instruction.
- Bloom’s high and low order questions – Teacher uses appropriate high-level and low-level questioning throughout the entire lesson cycle. The combination of the questions is aligned with the objective(s). Evidence of planning essential questions related to instructional objectives is present. Teacher consistently and appropriately matches Bloom’s levels to the objectives and/or standards.
- Prior knowledge – Teacher’s expectation of prior knowledge and student ability to scaffold learning from prior knowledge are appropriately aligned. Learning grain size is appropriate and effective for learning.

Moderately-high – Teacher generally uses questions effectively to scaffold instruction.

- Open ended questioning – Uses both closed and open questioning strategies, but generally uses open-ended questioning.
- Bloom’s high and low order questions – Teacher uses appropriate levels of high and low order questions during instruction as determined by the objective(s) most of the time. Some essential questions are generated and generally match Bloom’s level to the objective(s).
- Prior knowledge – Teachers’ expectation of prior knowledge and students’ ability to scaffold learning from prior knowledge are mostly aligned. Learning grain size is mostly appropriate for effective learning.

Moderately-low – Teacher rarely uses questions effectively to scaffold instruction.

- Open ended questioning – Some use of open-ended questioning is observed, but closed-ended questioning is the predominant questioning technique observed.
- Bloom’s high and low order questions – Relies on lower-order recall questions, but episodic use of higher-level questioning is observed. Teacher occasionally uses appropriate levels of questioning during instruction as determined by the objective(s).
- Prior knowledge – Teachers’ expectation of prior knowledge and student ability to scaffold learning from prior knowledge are somewhat aligned. Learning grain size is still too large for effective learning.

Low-Level – Teacher does not use questions effectively during instruction

- Open ended questioning – Almost exclusive use of closed-ended questions is observed.
- Bloom’s high and low order questions – Teacher did not use appropriate levels of questioning during instruction as determined by the objective(s). Focus on higher order thinking questions when only lower order was warranted or vice versa.
• *Prior knowledge* – Teachers’ expectation of prior knowledge and student ability to scaffold learning from prior knowledge are not aligned. Learning grain size is too large for effective learning.
Nature of Questioning Construct with Indicators

**High-level** – Teacher consistently and appropriately uses questioning effectively to diagnose problems with learning and improve instruction.
- Wait time – Teacher consistently uses appropriate wait time (3-5 seconds) to respond to students after a question has been asked.
- Follow up questioning – Teacher consistently and appropriately uses multiple questions to diagnose and/or scaffold instruction.

**Moderately-high** – Teacher generally uses questioning effectively to diagnose problems with learning and improve instruction.
- Wait time – Teacher uses appropriate wait-time (3-5 seconds) to respond to students after a question has been asked most of the time.
- Follow up questioning – Teacher uses questions 3 – 4 times during instruction to follow responses to their questions or students’ questions during instruction to diagnose and/or scaffold instruction. The follow-up questioning follows a logical progression from low-level to high-level in order to gain information from students, to stimulate thinking and appropriately scaffold learning.

**Moderately-low** – Teacher rarely uses questioning effectively to diagnose problems with learning and improve instruction.
- Wait time – Teacher occasionally uses appropriate wait-time (3-5 seconds) to respond to students after a question has been asked.
- Follow up questioning – Teacher use questions one or two times during instruction to follow responses to their questions or students’ questions during instruction to diagnose and/or scaffold instruction. However, the questions do not seem to have a significant contribution to the learning.

**Low-Level** – Teacher does not use questioning effectively during instruction
- Wait time – Teacher waits one second or less to respond to students after a question has been asked or answers their own questions.
- Follow up questioning – Teacher did not use questions to follow responses to their questions or students’ questions during instruction.
Self-evaluation Construct with Indicators

High-level – Teacher uses a variety (two or more) of strategy(s) to encourage students to self-evaluate in an effort to regulate and improve their own learning. These could include techniques such as the use of traffic lights, a self-assessment inventory, journaling, and/or a reflection statement.

- Teacher initiates self-evaluation - Teacher uses two or more self-assessment strategy(s) during instruction.
- Self-evaluation strategy(s) - The strategy(s) are clearly defined and relate to the objective(s).
- Student action - Students both understand how to assess their own learning and are able to use the strategies to assess their own learning. Students continually use self-assessment strategies throughout the lesson in order to maximize learning.

Moderately high – Evidence of one self-evaluation strategy is used during instruction in an effort to regulate and improve the student’s self-learning.

- Teacher initiates self-evaluation - Teacher uses one self-assessment strategy during instruction for improving student learning.
- Self-evaluation strategy(s) - The strategy the teacher uses is clearly described and relates to the objective(s).
- Student action - Students understand how to self-assess their own learning and use self-assessment strategies to assess their learning. This may be observed by a variety of student interactions such as student-student, student-teacher.

Moderately-low – Generic self-evaluation employed but the strategy is not explicitly tied to the regulation and improvement of self-learning.

- Teacher initiates self-evaluation - Teacher encourages students to evaluate their own learning during instruction, but in a very generic, unorganized, and/or not used for improving students’ learning.
- Self-evaluation strategy(s) - The strategy(s) for self-evaluation is not clearly described or generic.
- Student action - Students seem unsure of how to self-assess their own learning. This may be observed by a variety of student interactions such as student-student, student-teacher.

Low-level – Teacher does not use self-evaluation during instruction.

- Teacher initiates self-evaluation - The teacher did not explicitly encourage students to evaluate their own learning.
- Self-evaluation strategy(s) – The teacher does not provide strategy(s) for self-evaluation.
- Student action - Students may evaluate their own learning, however, the teacher does not use or is not aware that students are using self-evaluation in the classroom.
Student Affect Construct with Indicators

High-level – Teacher is sensitive to student affect, collects evidence through body language, facial expressions, and/or class work, and adjusts instruction accordingly. Teacher demonstrates expertise and polish in balancing content with student affect. Risk taking is actively encouraged and occurring frequently.

- Student behavior – Students are responsive to the teachers questioning and give meaningful answers that demonstrate engagement, critical thinking, and confidence.
- Teacher behavior - Teacher is continually sensitive to student feelings, body language, facial expressions, and/or class work.
- Teacher focus – The focus is learner-centered. Although sensitive to student affect, teacher demonstrates superb balance between teaching content and being sensitive to the affective needs of all students.
- Student interactions - A variety of student interactions are observed (i.e., student-to-student, student to group, student to teacher and teacher to student). Teacher actively and consistently encourages student interaction and creativity.

Moderately high – Teacher is sensitive to the affect of most students and shows general evidence of balancing teaching content with affect. Risk taking behavior is evidenced, but occurs as infrequent, episodic intervals.

- Student behavior – Students are responsive to questioning generally giving answers that demonstrate engagement.
- Teacher behavior – The teacher makes an effort to be sensitive to student feelings, body language, facial expressions, and/ or response to class work.
- Teacher focus – The focus is learner-centered based on the content. However, the teacher has not mastered how to reach all students in the class.
- Student interactions - There are frequent student and teacher interactions. Student creativity to voice ideas or suggestions.

Moderately low – Teacher shows limited sensitivity to student affect and tailors feedback for only a few students. Most emphasis is on teaching specific content. Although risk-taking behavior may be present, it is not encouraged.

- Student behavior - Students may give a token response to the teachers’ questions, but are not engaged in the lesson.
- Teacher behavior – The teacher makes some effort to be sensitive to a few students’ feelings, body language, facial expressions, and/or response to class work.
- Teacher focus – The focus is predominantly covering content, but the teacher only gauges the affect of a few students, usually the same 2-3 students.
- Student interactions – There are some student interactions with peers, but teacher is ambivalent toward student creativity to voice ideas or suggestions.

Low-level – Teacher does not attend to how feedback is received by the student. There is no evidence of consideration of affect. Total emphasis is on teaching specific content. Risk taking behavior is actively discouraged.
At low-level these indicators are associated with positive affect for most students:

- **Student behavior** - Students are unresponsive and uninterested in learning.
- **Teacher behavior** – The teacher makes no effort to be sensitive to student feelings, body language, facial expressions, or lack of response to class work.
- **Teacher focus** – The teacher’s focus is covering the content for the day’s lesson. Covering content refers to a content-centered lesson without attention to student affect about what is taught.
- **Student interactions** – The teacher actively discourages student interactions with peers and limits students’ creativity to voice ideas or suggestions.
Instructional Adjustments Construct with Indicators

High-level – Teacher consistently and effectively uses adjustments during instruction.
- Instructional strategies – Teacher makes appropriate and varied adjustments to instructional strategies based on evidence and feedback collected during instruction. Adjustment such as different strategies, providing context and models are consistently observed.
- Instructional timing – Teacher constantly and consistently uses just-in-time corrections to give appropriate feedback to students at the appropriate time.
- Instructional audience – Teacher maximizes instructional time and resources by addressing instructional adjustments to the appropriate audience. (i.e., individual, peer, group, or whole class).
- Grouping strategies – Intentional grouping practices are noted, and most students understand their role and participate in classroom activities.

Moderately high – Teacher predominately and effectively uses adjustments during instruction.
- Instructional strategies – Teacher generally makes adjustments to instructional strategies based on evidence and feedback collected during instruction. Adjustments such as using a different strategy, providing a context and using models are observed.
- Instructional timing – Teacher predominately uses just-in-time corrections to give appropriate feedback to students at the appropriate time.
- Instructional audience – Employs efficient use of instructional time and resources most of the time, and generally engages the appropriate audience for instructional adjustments. (i.e., individual, peer, group, or whole class).
- Grouping strategies – Intentional grouping practices are noted, but a few members do not understand their role or participate in classroom activities.

Moderately low – Teacher uses minimal adjustments during instruction.
- Instructional strategies – Teacher makes one or two adjustments during instruction. Adjustments are loosely tied to evidence collected during instruction. Teacher predominantly uses the direct teaching method.
- Instructional timing – Teacher uses a few just-in-time corrections to give feedback to students.
- Instructional audience – Employs efficient use of instructional time and resources some of the time, and occasionally engages the appropriate audience for instructional adjustments. (i.e., individual, peer, group, or whole class).
- Grouping strategies – Grouping practices are noted, but assignment of students to a group is random. Evidence of group role is absent or not clearly defined. Little group interaction related to lesson objectives is noted.

Low-level – No adjustments during instruction are observed.
- Instructional strategies – Teacher makes little or no adjustments to instruction and in general is non-responsive to student mastery of learning. Teacher predominantly uses the direct teaching method.
• Instructional timing – No just-in-time adjustments are observed. (A just in time adjustment is observed when a teacher allows students to struggle but steps in when limits have been met)
• Instructional audience – No consideration is given to the appropriate audience.
• Grouping strategies – No evidence of grouping or poor grouping practices is noted. Little or no group member interaction related to lesson objective is noted.
Evidence of Learning Construct with Indicators

High-level – There is evidence of learning for 75-100% of the students for one or more of the following indicators.

- All-student responses – Teacher consistently and appropriately collects and uses information from the class using all-student answering techniques. (e.g., Traffic lights, white boards, thumbs up).
- Individual responses – Teacher collects and uses information from almost all the students through verbal responses related to the lesson.
- Artifacts of learning – Most of the students complete a hybrid of traditional and authentic artifacts that demonstrates accuracy and understanding.

Moderately-high – There is evidence of learning for 50-74% of the students for one or more of the following indicators.

- All-student responses – Teacher occasionally collects and uses information from the class using all-student answering techniques (e.g., Traffic lights, white boards, thumbs up).
- Individual responses – Teacher collects information from over half the students through verbal responses related to the lesson.
- Artifacts of learning – At least half of the students complete a hybrid of traditional and authentic artifacts that demonstrates accuracy and understanding.

Moderately-low – There is episodic evidence that learning occurred. There is evidence of learning for 25-49% of the students for one or more of the following indicators.

- All-student responses – Teacher episodically uses all-students answering techniques to collect evidence of learning.
- Individual responses – Teacher episodically uses individual responses to collect evidence of learning.
- Artifacts of learning – Student artifacts consist of worksheet type assignments and teacher monitors only for accuracy and not for understanding.

Low-level – There is minimal evidence that learning occurred. There is evidence of learning for 0-24% of the students for one or more of the following indicators.

- All-student responses – Teacher never uses an all-student answering technique to collect evidence of learning.
- Individual responses – Teacher never uses individual responses to collect evidence of learning.
- Artifacts of learning – Student artifacts consist of worksheet type assignments, which are not reviewed for accuracy or understanding.
<table>
<thead>
<tr>
<th></th>
<th>Science Objectives</th>
<th>Science Question Quality</th>
<th>Science Nature of Questioning</th>
<th>Science Self-Evaluation</th>
<th>Science Student Affect</th>
<th>Science Instructional Adjustments</th>
<th>Science Evidence of Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATS1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATS2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATS3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATS4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATS5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATS6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS7</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>ATS8</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS9</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS10</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATS11</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS12</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ATS13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ATS14</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>ATS15</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATS16</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATS17</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ATS18</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS19</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ATS20</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATS21</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ATS22</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS23</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ATS24</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ATS25</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS26</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS27</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ATS28</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS29</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS30</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATS31</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATS32</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATS33</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>ATS34</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: ATS1 signifies AssessToday science observation number one
APPENDIX D

MATHEMATICS OBSERVATION DATA
<table>
<thead>
<tr>
<th></th>
<th>Objectives</th>
<th>Question Quality</th>
<th>Nature of Questioning</th>
<th>Self-Evaluation</th>
<th>Student Affect</th>
<th>Instructional Adjustments</th>
<th>Evidence of Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATM1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ATM3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>ATM4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM5</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>ATM6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>ATM7</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM8</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM9</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>ATM10</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM11</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM12</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM13</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM14</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM15</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM16</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM17</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM18</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM19</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ATM20</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM21</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ATM22</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM23</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM24</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>ATM25</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ATM26</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM27</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM28</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM29</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ATM30</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ATM31</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM32</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM33</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ATM34</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ATM35</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM36</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM37</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM38</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM39</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM40</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM41</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM42</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>-------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>ATM43</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM44</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM45</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>ATM46</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ATM47</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ATM48</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM49</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM50</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM51</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM52</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM53</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM54</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM55</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM56</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM57</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ATM58</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM59</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM60</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM61</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATM62</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM63</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ATM64</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: ATM1 signifies AssessToday Mathematics observation number 1
APPENDIX E

SCORES FOR OBSERVATION PAIRS
<table>
<thead>
<tr>
<th>Observation Pair</th>
<th>Observers (x,y)</th>
<th>Objectives</th>
<th>Question Quality</th>
<th>Nature of Questioning</th>
<th>Self-Evaluation</th>
<th>Student Affect</th>
<th>Instructional Adjustment</th>
<th>Evidence of Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(M1,M2)</td>
<td>(0,0)</td>
<td>(1,1)</td>
<td>(1,2)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,2)</td>
<td>(1,2)</td>
</tr>
<tr>
<td>2</td>
<td>(M1,M2)</td>
<td>(1,1)</td>
<td>(2,3)</td>
<td>(2,3)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
</tr>
<tr>
<td>3</td>
<td>(M1,M3)</td>
<td>(0,0)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(2,3)</td>
</tr>
<tr>
<td>4</td>
<td>(M4,M5)</td>
<td>(0,0)</td>
<td>(1,1)</td>
<td>(1,2)</td>
<td>(0,0)</td>
<td>(1,2)</td>
<td>(1,1)</td>
<td>(1,1)</td>
</tr>
<tr>
<td>5</td>
<td>(M4,M5)</td>
<td>(2,1)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(1,1)</td>
<td>(2,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
</tr>
<tr>
<td>6</td>
<td>(M4,M5)</td>
<td>(0,0)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0,0)</td>
<td>(1,1)</td>
<td>(0,0)</td>
<td>(1,1)</td>
</tr>
<tr>
<td>7</td>
<td>(M4,M5)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0,1)</td>
<td>(0,0)</td>
<td>(1,0)</td>
<td>(0,0)</td>
<td>(1,1)</td>
</tr>
<tr>
<td>8</td>
<td>(M4,M5)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,0)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
</tr>
<tr>
<td>9</td>
<td>(M4,M5)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(0,1)</td>
<td>(1,1)</td>
</tr>
<tr>
<td>10</td>
<td>(M4,M5)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(2,1)</td>
<td>(0,0)</td>
<td>(2,1)</td>
<td>(2,1)</td>
</tr>
<tr>
<td>11</td>
<td>(M4,M5)</td>
<td>(2,1)</td>
<td>(2,1)</td>
<td>(1,2)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(1,2)</td>
<td>(1,2)</td>
</tr>
<tr>
<td>12</td>
<td>(M4,M5)</td>
<td>(0,0)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(1,1)</td>
<td>(2,2)</td>
<td>(2,2)</td>
</tr>
<tr>
<td>13</td>
<td>(M4,M5)</td>
<td>(0,0)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
</tr>
<tr>
<td>14</td>
<td>(M4,M5)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(3,3)</td>
<td>(2,2)</td>
<td>(3,3)</td>
<td>(2,2)</td>
<td>(2,2)</td>
</tr>
<tr>
<td>15</td>
<td>(M4,M5)</td>
<td>(1,1)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(2,2)</td>
</tr>
<tr>
<td>16</td>
<td>(M4,M5)</td>
<td>(1,1)</td>
<td>(1,1)</td>
<td>(2,2)</td>
<td>(2,2)</td>
<td>(1,1)</td>
<td>(1,2)</td>
<td>(1,2)</td>
</tr>
</tbody>
</table>

Note: M1 represents math observer 1, S1 represents science observer 1. The corresponding ordered pairs are the scores the corresponding observer assigned for the construct.
REFERENCES


doi:10.1037/h0037632


166


