ASSESSING THE VISUAL ARTS:

VALID, RELIABLE, AND ENGAGING STRATEGIES

by

Alexandria English

A Project Submitted to the Faculty of
The Evergreen State College
In Partial Fulfillment of the Requirements
For the Degree
Master in Teaching
2010

This Project for the Master in Teaching Degree

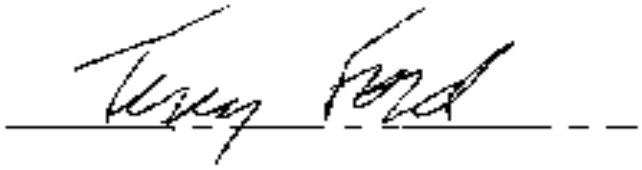
by

Alexandria English

Has been approved for

The Evergreen State College

by



Dr. Terry Ford

June 2010

ACKNOWLEDGEMENTS

I would like to thank everyone that has helped me on this long journey. An extra thanks to Dr. Terry Ford for reading pages and pages of rough drafts and revisions to get me to this point. Also, thanks to all the faculty members of the MiT program for their dedication and passion for their students.

I would especially like to thank my family for their never ending belief and support. Thank you to my fellow cohort members for keeping me focused, motivated, and sane. And of course, thank you, Christopher, for washing dishes and cleaning the house when I was too tired, being my shoulder to cry on, forcing me to have fun when I needed it, and letting me sleep in on those rare weekend mornings.

ABSTRACT

This paper attempts to answer the question what are alternative assessment methods that teachers and students find reliable and valid. To answer this question, this paper discusses the history of art education and examines research that investigated portfolio assessment, self and peer assessment, performance assessment, and teacher and student perceptions of alternative assessment.

While there is still more research to be done in the area of visual arts assessment, some conclusions can be drawn from the research investigated here. Portfolio assessment, self and peer assessment, and performance assessment were all found to be reliable methods of assessing students. However, there is more to be discovered in regard to teacher instruction practices and curriculum development. There are also implications for teachers and instructors that plan to use alternative assessment in their classrooms.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION

Introduction

"Significant evidence..."

"State content standards..."

"Assessment..."

"Learning outcomes..."

These terms all have the ability to elicit a certain feeling, a sense of numbness or hopelessness from art educators. This reaction is not unique. Many teachers believe that visual art assessment has very little connection with the problems of classroom instruction, classroom curriculum, and classroom management (Beattie, 1997). But it does.

Assessment should not be a meaningless task that requires hours of time form teachers. It is a vital part of quality teaching and better instruction. Good classroom practice involves pre-assessing for prior knowledge; building on that prior knowledge through instruction; re-assessing; re-teaching based on assessment findings; and a final assessment. This is an ongoing process that runs concurrent with instruction for the purpose of learning (Beattie, 1997). This includes learning in the visual art classroom. In order to understand student knowledge, learning, and processes, visual art teachers must provide various valid means of assessment to better understand their students thought processes in the arts as well as in other content areas.

Research has shown that most teachers spend nearly one half of their work time doing assessment related work (Conklin & Stiggins, 1992). That is quite some time. Understanding assessment and knowing and using effective strategies and procedures can give art educators the tools they need to properly utilize this time and make educated decisions instead of blind guesses. This project explores effective assessment strategies in the visual arts that are valid, reliable, and engaging for students as well as teachers. These strategies include both alternative

and standardized forms of assessment. By researching, reviewing, and analyzing these strategies, I can begin to create an effective art assessment program that allows me to continuously discover and monitor student strengths, weaknesses, and progress; improve instructional strategies based on students' learning; and use different information from students to effectively manage the classroom (Beattie, 1997).

Rationale

The visual arts can be a gateway to opening a student's soul, give a student a voice, and act as a tool in critically examining the world. But it can also be a source of frustration, separation, and fright. There is a choice that happens in a student's mind when asked to participate in an art activity. He or she can either join in the activity or reject it with the mentality that they are not an artist; they have no talent; they do not have the right personality, background, or knowledge. Now, that choice has been made for the student. Due to the No Child Left Behind Act (NCLBA) as well as Washington state Essential Academic Learning Requirements (EALR) and Grade Level Expectations (GLE), students are required to have an education in the visual arts. Students are also required to show proficiency in these areas based on state performance standards. In recent years, there has been a growing dissatisfaction with the United States' traditional methods of standardized, multiple choice testing. The result is a growing interest in different forms of alternative assessment at different levels. Educators are discussing portfolios, exhibits, writing and hands-on experiments as new, more effective means of assessment in classroom curriculum (Aschbacher, 1992).

The idea of assessment has been unwelcome and unaddressed in arts education for several reasons. According to Eisner (1966), there are five reasons for this. One, visual art assessment is dependent on judgments of the quality and craftsmanship of student work. Some

educators see these judgments as blocks in student potential. Judgments can stifle creativity and assessments some art educators see assessments as value judgments. Second, both assessment and evaluation include some form of measurement of student performance. Some art educators believe that the arts should focus on the experience of creating art and that experience cannot be quantified, leaving measurement and the arts incompatible. Third, assessment and evaluation have traditionally been related to the final product or outcomes of the student's efforts. Most art educators regard the process that students engage in to be what is important, not the product. Fourth, many art educators believe that the art field and the desired objectives of art education are anything but standardized. Assessment has been traditionally seen as closely associated with testing and particularly standardized testing. And lastly, assessment and evaluation are closely linked to grading. Some art educators believe that grading, especially at a young age is worse than irrelevant. It is actually considered harmful to the student.

So why even look at assessment? Why bother with the struggles that so many other art educators have fought? As Aschbacher (1992) pointed out, assessment can serve needs at all levels of the education hierarchy; for example, assessment helps educators set standards, create instruction pathways, motivate performance, provide diagnostic feedback, assess and evaluate progress, and communicate progress to others.

In order to find effective visual art assessments, I will examine research that investigates different assessment strategies, the validity and reliability of both alternative and standardized assessment methods, effects on student learning, and student perceptions of different assessment methods. This research investigates assessment methods from various academic content areas as well as performance activities such as physical education from different parts of the United States and different countries, especially the United Kingdom. I will analyze and

compare sets of studies relating assessment methods and educational systems in other countries to assessment methods and educational systems in the United States.

Definition of Terms

To build a framework for analyzing visual art assessment, certain terms that are used in the professional literature and research should be defined to create a level ground for all discussions in the following chapters.

Assessment is an ongoing process aimed at understanding and improving student learning. It involves making teacher expectations explicit and public; setting appropriate criteria and high standards for learning quality; systematically gathering, analyzing, and interpreting evidence to determine how well performance matches those expectations and standards; and using the resulting information to document, explain, and improve performance. When it is embedded effectively within larger institutional systems, assessment can help us focus our collective attention, examine our assumptions, and create a shared academic culture dedicated to assuring and improving the quality of education (Angelo, 1995). Angelo (1995) discovered five themes that are involved in assessment: assessment should focus on improving student learning - the focus of assessment should not be limited to the classroom, but include the wide range of processes that influence learning such as critical thinking and problem solving; assessment is a process embedded within larger systems; assessment should focus collective attention and create links and enhance coherence within and across the curriculum; and tension between assessment for improvement and assessment for accountability must be managed. Assessment generally refers to the appraisal of individual student performance, often but not necessarily on tests. This differs from evaluation. For the purposes of this project, assessment can be simply defined as the method or process used for gathering information about people,

programs, or objects for the purpose of evaluation.

Evaluation includes judgments of value concerning the worth of any aspect of the educational system in play including student learning, teacher effectiveness, program, quality, and educational policy. Evaluation generally refers to the appraisal of the program - its content, the activities it uses to engage students, and the ways it develops thinking skills (Eisner, 1966).

Standardized testing includes tests that are administered and scored in a consistent manner. The tests are designed in such a way that the questions, conditions for administering, scoring procedures, and interpretations are consistent and they are scored and administered in a predetermined, consistent standard. I will examine two different types of standardized testing: normative referenced and criterion referenced. *Normative referenced or norm-referenced* includes testing that compares one test taker to his or her peers. This testing yields an estimate of the score of the tested individual. This estimate is derived from the analysis of all test scores and possibly other relevant data concerning the population tested. This differs from criterion referenced testing. *Criterion referenced testing* includes testing that translates the test scores into a statement about a student's learning or knowledge. The objective of these tests is simply to see what material the student has learned.

Alternative assessment is assessment that deviates from the traditional paper and pencil methods. In the education industry, alternative assessment is in direct contrast to what is known as performance evaluation, traditional assessment, standardized assessment or summative assessment. Alternative assessment is also known under various other terms including: authentic assessment, integrative assessment, holistic assessment, assessment for learning, and formative assessment. Alternative assessments are used to encourage student involvement in their assessment, their interaction with other students, teachers, parents, and the larger community. Some characteristics of alternative assessment include: students

perform, create, produce, or do something; students tap higher level thinking and problem solving skills; students use tasks that represent meaningful instructional activities and learning opportunities; activities invoke real world application; people, not machines, do the scoring utilizing human judgment; and it requires new instructional and assessment roles for teachers (Beattie, 1997).

Formats vary. Beattie (1997) stated that such different formats - portfolios, journals, sketchbooks, diaries, integrated performances, group discussions, exhibitions, and multimedia are all considered different forms of alternative assessments in the visual arts.

Authentic assessment utilizes realistic, meaningful, open ended problems true and specific to a certain discipline

Portfolio assessment, in the past, has been seen as a collection of student artworks or products. More recently, portfolios are being used as a means of understanding and revealing processes. It can now be defined as a purposeful collection of student work that tells the story of the student's efforts, progress, or achievement in a given area. Portfolios may be assessed in a variety of ways. Each piece may be individually scored or the portfolio may be assessed merely for the presence of required pieces, or a holistic scoring process might be used and an evaluation made on the basis of an overall impression of the students collected work. It is common that assessors work together to establish consensus of standards or to ensure greater reliability in evaluating student work. Established criteria are often used by reviewers and students involved in the process of evaluating progress and achievement of objectives (Beattie, 1997).

Classroom Based Performance Assessments (CBPAs) are built from Washington State's learning standards. State curriculum specialists create tasks and questions that model good assessment methods and provide these to local school districts. As the name suggests, these

assessments are given in the classroom by an instructor. Classroom based assessments and classroom based performance assessments are being used to inform the teacher and the state if students are gaining key skills and knowledge in social studies, the arts, and health/physical fitness. By the end of the 2008-09 school year, school districts hoped to have in place in elementary, middle, and high school assessment strategies to assure that students have an opportunity to learn the essential academic learning requirements (EALRs) in social studies, the arts, and health/fitness. Currently, the visual art CBPAs focus on the elements and principles of art and design. Beginning with the 2008-09 school year, districts will annually submit an implementation verification report to the Office of the Superintendent of Public Instruction (www.k12.wa.us).

Essential academic learning requirements (EALRs) are Washington state learning standards that provide an overview of what students should know and be able to do in grades K-12. Generally, they include the following: read with comprehension, write effectively, and communicate successfully in a variety of ways and settings and with a variety of audiences; know and apply the core concepts and principles of mathematics; social, physical, and life sciences; civics and history, including different cultures and participation in representative government; geography; the arts and health and physical fitness; think analytically, logically, and creatively to integrate different experiences and knowledge to form reasoned judgments and solve problems; and understand the importance of work and finance and how performance, effort, and decisions directly affect future career and educational opportunities (www.k12.wa.us).

Grade level expectations (GLEs) are detailed standards that state what students should know and be able to do by each grade level. GLEs are aligned from kindergarten through grade 12 so that parents, students, and educators can see how skills and knowledge build from year to year.

GLEs are developed for each content area (www.k12.wa.us).

Reliability refers to the degree to which an evaluation score is consistent across time, judges, or forms of assessment. In relation to time, students can expect the same grade with the same level of study and preparation despite the day, the time, or the school schedule. The judge or instructor has the same standards. And the assessment form should be consistent. For example, if a student is taking a make-up exam, the exam should have the same content based on the same learning objectives and the same weight of the course grade as the original exam (Beattie, 1997)

Validity refers to the degree to which a score is meaningful and appropriate for its intended purpose. This means that validity refers to whether and to what degree a score means what the instructor thinks or says it means. Validity is a characteristic of a score put to a particular use, not a characteristic of a test or assessment in itself (Beattie, 1997).

Controversies

Controversy exists in the visual arts in all possible areas of education - assessment, curriculum, accountability, and purpose. For this paper, I will be focusing on the controversy involving instructional practice and assessment methods in the visual arts, namely how to assess the arts. There has been long standing debate on assessment in generally and now there is even more focus on assessment in the visual arts as the arts are making a comeback in certain state education standards.

There are many differing opinions on how the visual arts should be taught, each offering a different opinion on art assessment and evaluation. The Discipline Based Art Education stance views the visual arts as important in standing alone, separate from integration into the other content areas. This viewpoint places the visual arts as an accountable, purposeful content area

that can be viewed as a serious discipline by the nation and by state districts. In the discipline based classroom, students are asked to create and manipulate images in the fashion of adult artists. Students are seen as eventual adults who will need to learn the design principles and elements of art to create a conceptual work of art (Rush, 1987). There is also the opinion that the arts are more beneficial when integrated into the other content areas as the visual arts can aid students in better understanding abstract concepts and skills. Some educators have voiced that DBAE has a narrow vision and endangers student imagination, creativity, and freedom (London, 1988; Lederman, 1988; and Lidston, 1988). Controversy has been bubbling regarding the philosophical and fundamental practices in the visual art field. The cognitive aspects of visual art education are being questioned for validation and acceptance in the public school system. In 1988, the National Endowment for the Arts published *Toward Civilization: A Report on Arts Education*. This study took arts education further - incorporating governance, education, arts and business into the classroom, much like the trend of bringing business into schools. The report also encouraged that school systems create curricular and instructional assessment models and advocate for high standards for state and local art curriculum. According to Topping (1990), the field of art education is being shaken by different opinions of art education, from arts in general education to interdisciplinary education. National, state and local departments of education are all trying to clarify the mission and priorities of an art education. There is a need to reach an agreement on what students should learn in an art classroom about art. The Getty Center for Education in the arts and the National Endowment for the Arts both advocate for a curriculum that provides art production, art history, aesthetics, art criticism, and knowledge regarding art in civilization (Topping, 1990). Some argue that there should be a reform of the studio approach to teaching and learning art, where the student's ideas and feelings should remain the focus of the art curriculum. The studio approach requires structure and sequence

and demands reflection, analysis, and synthesis. However, these are not practiced by all art teachers. Perhaps art teachers and educators are not giving these areas adequate attention.

This seems in contrast to the standards and objectives that the national and state education departments have developed for the visual arts. Teachers and students are given benchmarks and standards as goals for the curriculum. Topping (1990) suggests that assessment and evaluation should be used to help teachers improve their instructional programs and practices. Testing rarely works to help teachers reform curriculum.

With the push of standardized testing, there is also a controversy surrounding the methods to assess the visual arts. Art education has been missing a focused standardized testing force. As a result, art education has escaped many of the problems prevalent in other academic areas (Gruber and Hobbs, 2002). Lowenfeld (1982) believed that progress indicated learning and that progress could be measured. In recent years, there has been a surge of criterion referenced standardized assessment developed for various performance based content areas including the visual arts, social studies, and health and physical education. There are differing opinions on exactly how to assess a performance based on a norm or a standard, especially when that performance is so strongly tied to personal creativity and expression. With the new call for accountability in education and assessment of all academic areas, including art education, assessment has come to the forefront. Art educators should be doing their homework (Gruber and Hobbs, 2002).

Limitations

I had hoped to limit my research to studies that directly focused on visual art curriculum and assessment methods at the high school level in the United States. Upon reviewing the available research, that hope was obliterated. There are limited studies that pertain to alternative assessment in the visual arts. The research studies included in this paper vary in grade level, content area, and geographic location. However, the studies do focus on one common theme, the use of alternative assessment as a means to evaluate a performance task.

Summary

This project explores effective assessment strategies in the visual arts that both students and teachers can find valid, reliable, and engaging. I will analyze both alternative and standardized forms of assessment used by different educators in different content areas. By researching, reviewing, and analyzing these strategies, I can create a collection of different assessment methods that can improve my instructional techniques, curriculum, and classroom management.

In the next chapter, I will be analyzing the history of art education and its influence on student assessment in the art classroom. Different popular beliefs, art education movements, and political agendas have influenced art education and art assessment for better and for worse. This has led to different forms of visual art assessment in the classroom. In Chapter Three, I will critically examine research studies and literature on the development of alternative assessment in a visual art curriculum that is valid and reliable. And finally, in Chapter Four, I will discuss the findings of the research of Chapter Three. I will discuss possible answers and strategies that visual art teachers can use as assessment methods in a visual art classroom. These strategies may help teachers to create a curriculum that is better suited for the individual student and the whole class.

CHAPTER TWO: HISTORICAL BACKGROUND

Introduction

In the previous chapter, the question of what effective strategies for visual art assessment can be valid, reliable, and engaging was introduced, including the reasons for its importance as well as the reasons that it has not been explored in depth, and the controversies surrounding visual art assessment. This chapter briefly focuses on the history of art education in the United States, the beginnings of assessment in the public school system, and the reasoning for such assessments and curriculum in the public school system. This chapter will serve as a background for Chapter Three, the review of the research.

The Beginnings of Art Education

Historically, the beginning of art education lies in Greece. In Greek antiquity, different philosophers including Aristotle (1941), believed that the standard branches of education were reading and writing, gymnastic exercises, music, and drawing. Of these, reading and writing and drawing were regarded as useful for the purposes of life in a variety of ways. Gymnastic exercises were thought to instill courage in an individual.

It is also with Greece that we see where our contemporary ideas of education stem and how they came to be. Aristotle did state that drawing is regarded in the same light as reading and writing. Dewey (1925) saw it another way. Modern thought accepts the Greek view that knowledge is contemplation and that can be revealed naturally. In regard to the arts, Greek society accepted the useful arts as modes of practice, but it also rejected the fine arts. Greek society viewed "contemplation" as the areas of science and logic and "production" as being the

art worthy of exploration. Even with drawing regarded as useful, it still came second to thought. This brings us to the contemporary views of art and art education. Dewey (1925) believed that the contemporary theories of art education are very inconsistent because they are only, in part, interpretations of the art movement of a particular time period and, in part, interpretations of Greek opinions.

This thinking was similar to the ideas of Spring (2005). Spring (2005) stated that the common school, what is now referred to as the public school, suited the interests of the time and the people. The history of public education has shown that education is dependent upon the predominant belief of what is important to the people at any particular time period. Initially, the belief was religion and it has evolved over the years into Americanization, acculturation, and now, human capital. Including the visual arts into the public education system is no different. It began with a belief and a need - a belief that students were human capital and a need to teach drawing for an industrial age.

The Industrial Age

Art education and instruction was advocated as early as 1770 by Benjamin Franklin. The official beginning of art instruction in American began in the nineteenth century. Art was taught in both private and public schools where individual teachers nominated to teach it. There were no national organizations of art teachers, no state laws requiring that art be taught, and art instruction was based on the personal views and opinions of the individual teacher (Eisner,1966).

By the 1860's, the nation was bustling with industrialization, booming with new technologies as a result of the technological and mechanical advances of the Civil War. Massachusetts was the most industrialized state of the nation. Not only was it the most industrialized, it also sported the most fully developed public school system. This was prime ground for creating a large, skilled work force from students. Due to the economic situation, art became significant. It took on a new role in society and was considered an industry service and an important vocational skill (Eisner, 1966). During this time, Massachusetts led the way for educational innovations including attendance, school inspections, required school reports and teaching drawing in local schools. This went on to serve the industrial need. Art education became a means to develop human capital. Large cities needed a workforce that could work with their hands, provide quality craftsmanship, and diagram and draw engineering plans and blueprints. In 1870, the subject of drawing was mandated by law as a school subject (Efland, 1990). Art education had officially made its way into the public school system. It would go through several transformations with each international art movement and national educational reform.

The Turn of the 20th Century

In the time between the late 1800s and the First World War, art education transitioned from technical drawing to a more inclusive education. Art education was still heavily influenced by business and social science. Several schools began to pop up in highly populated, industrialized areas offering drawing and design courses as well as the technical arts. Art and craft clubs began to appear at this time as well as the development of the arts and crafts movement in art. This gave women a chance to gain an art education in craft while men were already taking courses in drawing and design. This movement also redefined how administrators

and the public viewed art education. Administrators argued that the arts and crafts were really complex, complicated processes that required students to problem solve for construction, design, and artistic craftsmanship. These processes and thinking were the foundation of the modern industry. This differed from the opinion that arts and crafts were just busy work for students. Teachers were expected to demonstrate a business mindset while also being knowledgeable about art techniques and media, and encouraging student interest and problem solving.

In 1916, The Commissioner of Education, David Snedden, addressed the problems in art education at the Eastern Arts Association conference in Springfield, MA. Snedden stated that the art education of the past fifty years had made no difference in state industry or the state work force. Snedden did not agree with the long standing goal that art education was to prepare students to be future workers. He noted that industrialization and the growth of the market had reduced the number of arts related skills and needs in jobs. Instead, Snedden proposed that art education be integrated into art courses so that all students could have access and experience in art activities. He also stated that these art activities should be natural and neutral, not predetermining and molding students to a desired role. This created a new curriculum and a different method of teaching art. These different classes can be seen in the articles of the *School Arts Book*, a publication that ran from 1903 to 1913. These classes included: Appreciation of Beauty, Nature Drawing, Handicrafts, Model and Object Drawing, and Color and Design (Efland, 1990).

Progressive education began in the work of Francis Wayland Parker and John Dewey at the turn of the 20th century and into the mid 1900s. They studied children's natural interests as a base for curriculum. From this point, they gave life to the current curriculum with nature study

and manual occupations. Both Parker and Dewey recognized the importance of art for stimulating a child's observations and interests (Efland, 1990). In 1938, John Dewey wrote *Experience and Education*. With this text came a shift in art education. Art education went from a practice occupied with proper drawing, art study, and hand eye coordination to a practice focused on unlocking the creative potential of children (Eisner, 1966). Despite the slow advances in art education at this time, there were still great strides being made as seen from the rise of large cities hiring art specialists and supervisors to work with teachers in the schools to the rise of professional associations for art and manual training teachers.

The Great Depression and the First World War

Due to the advancements in educational theory proposed by John Dewey's concept of experience as the reconstruction of knowledge, so too was there a reconstruction of social institutions. According to Efland (1990) and Dewey (1925), art was seen as an experience, as a way to construct knowledge and understanding. Art became more about personal expression, and art became a means to change individuals and society. Others began to see art as a way to solve problems in the home, in the school, and in the community. Art was thus seen as an integral part of human life and activity. The different studies of art, such as art history, came to show how art was linked to worship, the state, the craft, and personal expression. In this light, it made sense to re-constructivist educators to integrate art into the rest of the curriculum. Eisner (1990) believed that this new concern with children's creativity resulted in an interest in integrating the arts into other content areas in the classroom, an interest in the relationships of different content areas to each other and to the arts. Schools and educators were concerned with meaningful learning and built programs and curriculums around problems and projects. This switch from studying art as a part of the greater curriculum as opposed to individually could

also be a result of the economic stress of the Great Depression.

Before the crash of 1929, art had a great standing and a bright future in the world of education. Royal Bailey Farnum stated, "Art education in the United States has never been on a firmer footing that at the present time. It faces a future secure in the knowledge that during the past ten years its social, economic, and educational values have been demonstrated and acknowledged and generally put into practice (Efland, 1990, 78)." He continued stating that more powerful work was ahead for art education. That prediction would not be fulfilled, however, as the United States plummeted into the worst economic crisis in its history. As a result, education was hit hard. More than 2,280,000 children did not attend schools because the schools were forced to close. This was mainly in rural parts of the country where 2000 schools in 24 different states did not open in 1933. Many schools also shortened the length of their school terms and decreased teacher salaries. In 1933, the US Department of Education conducted a survey of 700 typical cities to see what the effects of the Depression on art, music, physical education and the industrial arts. Thirty-five cities had eliminated art programs entirely, and sixty-seven had reduced art instruction. However, we are not sure how many schools had previously taught on a regular basis which could affect the statistics (Efland, 1990). This moment would begin a continuing trend in art education for many years to come.

The Second World War to the Present Day

During the war, art education was under a struggle to stay afloat. Art education was seen by some as a means of preserving and defending democracy, as demonstrated by various Victory Posters that supported the war movement. For other art educators, the war presented a great challenge. Educators had to show that art was ideologically committed to the fight to preserve freedom and democracy, the same freedom that promoted and protected artistic self-

expression. The debate of free speech and censorship began. While art was considered a means of self expression and free speech, art educators were worried that art education would also be seen as a means of escape and national disloyalty (Efland,1990). Political freedom required giving up some artistic expression.

It was also during this time that Viktor Lowenfeld published the first edition of his book *Creative and Mental Growth*. Lowenfeld (1982) viewed the goal of art education as not the art itself or the aesthetic product or the aesthetic experience, but rather the child who grows up more creatively and sensitively and applies his experience in the arts to whatever life situations may be applicable. A new type of art education was introduced to the public schools that integrated the whole child through art. Art became important because it enabled creative problem solving skills to develop long before they could develop in other areas of education. The purpose of the arts was to develop creativity with the hope that this creativity would transfer to other educational areas and human activities as creativity was seen as a part of everyday human activity. Efland (1990) described how art activities geared toward special interest topics of children - such as home, family, friends, vacations, and hobbies - invited students to express their lives and learning through the things they created. According to Lowenfeld (1982), when children expressed themselves through art, they showed their knowledge of the environment. By evaluating the children's stages of creative development, educators gained a greater understanding of the children and their needs. Here we see one of the first moments of using assessment and evaluation to gain insight into the student as well as the student constructed art products helping children develop inquiry and seek answers about their environment to make new connections and form new concepts. However, there have been critiques of this approach to art education in regard to assessment and teacher role.

Smith (1984) stated that this method of art education does not treat art as a serious discipline but rather as a free flowing approach that allows the students to create independently. This type of education did not lend itself very well to assessment. Lowenfeld (1982) agreed to this critique because of the difference in thinking required for creativity. Creativity fosters divergent thinking involving many different solutions while the assessment at the time, mainly quantitative tests, focused on convergent thinking with only one possible answer.

Also during this time, two university professors were attempting to understand what makes a work of art successful and what makes the process of making art successful. Aurthor Wesley Dow, professor of Fine Arts at Columbia University, was concerned with the principles incorporated into a successful work of art. He came to the conclusion that a work of art was successful if it possessed composition. Composition was composed of three elements - line, value, and color; and five principles - opposition, transition, subordination, repetition, and symmetry. These elements and principles would change over time but would also become the standard for assessing student work. Walter Sargent, professor of Aesthetic and Industrial Education at the University of Chicago, was concerned with the psychology of children's art and the methods children use to produce art. Sargent focused on the process of creating art, stating that drawing was a visual language - a way to produce, form and develop ideas (Eisner, 1966).

The Cold War did not help arts education. According to Spring (2005) the Cold War generated a wealth of educational policies designed to use schools to strengthen national defense. More emphasis was placed on the content areas of math and science; and, as a result, funding was provided to the content areas of math and science even though education panels warned against the imbalance in federal funding and curriculum reform. Through the cooperative research act of 1954, educators sought to improve education through basic

research and testing. So began a trend that the United States believed it was failing economically at the global level because of its educational system. This act further pulled the focus on education away from art education and pushed focus toward the math and sciences.

There was still some hope for art education to be kept alive however. Some people were voicing their opinions and misgivings. Two of those people were Manuel Barkan and Harry Broudy. In 1965, the US Office of Education funded the Pennsylvania State University Seminar, where art educators came together to deliberate the benefits of a structured, discipline based approach to art education. "It was a unique gathering of the clans, in which the influence of Barkan and his optimism for the future was unmistakable. It was also the first conference I know of in which the idea of studying art in order to learn about it was introduced to the field since early in the century" (Lanier, 1991, 24). As hopeful as the Penn State Seminar seemed, the sixties also saw a decrease in funding for art education in public schools. According to Lanier (1991), as the 1960s came to a close, funding for the arts decreased to practically nothing. What funding was left at the end of the decade was reassigned to the National Endowment for the Arts for artist-in-school programs. Twenty years after this seminar, the term Discipline Based Art Education (DBAE) was used to describe this art education reform (Dobbs, 1988). Inspired by the ideas of Barkan and Harry, this style of art education differed from the previous ideas of Viktor Lowenfeld in that this art education focused on creative self expression and the idea that art can be treated as a serious content area. DBAE includes four areas of instruction relating to art - aesthetics, studio art, art history, and art criticism. DBAE also differed from other types of art instruction in that it focused on the inherent value of studying art (as opposed to how art can help students better understand other content areas), the skills needed to make art, and the order of creating and presenting the subject matter in the student art work. According to Greer (1984), DBAE is a much better type of art instruction for two very important reasons - if art is

treated as a serious discipline, it will be viewed as a serious discipline by national and state school boards as well as the general public, increasing art education support; and teachers and students can be better and more easily assessed on the art products, processes, and performance since Discipline Based Art Education stresses content.

There were some critiques of Discipline Based Art Education. Vincent Lanier (1991), one of the first participants in the DBAE program with the John Paul Getty Center for the Education in the Arts, believed that DBAE offered only the studio arts as a means to represent ideas. Lanier noted that there was no accountability required of students and teachers since there were no checks or balances or standards. He also noted that many educators agreed with Discipline Based Art Education, stating that they had been using such a curriculum in their own classrooms. However, there was little evidence of art history or criticism in school district reports.

Students as Human Capital

In the 1970s, the accountability movement spread and states and local communities began to require that schools publish achievement test scores annually. These test scores were used to measure the schools' success and kept the power in the hands of the educational experts. This resulted in students taking more and more achievement tests to satisfy the requirements of accountability. This emphasis on achievement testing became the central feature of the national educational policy in the early 1990s, creating national and state standardized achievement tests (Spring, 2005).

The 1980s saw an increase in closer ties between big business and public education. We can begin to see the federal government view students in the school system as human capital in need of fixing to build a better economy. Two reports can be seen as responsible for this new

viewpoint. In 1983, the publication *Nation at Risk*, blamed the nation's public schools for America's failure to compete with other countries in the world market, particularly Japan and West Germany (Spring, 2005). *Action for Excellence*, the second report released in 1983 by the Task Force in Education for Economic Growth, advocated closer ties between big businesses and public schools. The introduction to the report stated that the Task Force and the federal government believed that businesses should be more involved in education because they hold the role of employer (Spring, 2005). The Task Force believed that if businesses became more involved in the design and method of education and curriculum, America would be more competitive as an economy, more socially stable, and promote the national defense, well-being, and prosperity of the country (Spring, 2005). With the election of George H.W. Bush, the emphasis on public schools contributing to the economy continued. On April 18, 1991, Bush revealed Goals 2000, a plan for achieving national education goals and standards by the year 2000. The Bush administration proposed creating voluntary American Achievement Tests for grades 4, 8, and 12. The tests would cover five core subjects, and students would be measured by world class standards. To accomplish this goal, the Bush administration, in cooperation with Congress and the National Governors Association, created the National Council on Education Standards and Testing. Assessment was officially standardized (Spring, 2005). The Clinton administration kept the trend of standardized testing and student preparation for the work force by funding school programs that combined school-based and work-based learning to prepare students to become part of the American workforce. In 1991, the Clinton administration passed the School to Work Opportunities Act which funded programs that combined education and employment. The legislation provided funding that created and supported vocational classes and career counseling. These programs provided on the job training and paid work experience (Spring, 2005). This was a major step in viewing students as human capital.

During this same time period, The National Endowment for the Arts released the report *Understanding How the Arts Contribute to Excellent Education* (Spring, 2005). This report stated that the arts have been used as a means for gaining better academic performance and student understanding. Art curricula that have used this style of art education have been doing so since the 1970s, but this curriculum is just now currently coming into its own. This method of art instruction was termed interdisciplinary education. This theory of education used art activities to help students in the subject areas of math, reading, and science. While integrating art into different content area curriculum, it was noted that this type of education was not to be a substitute for art education independent from other areas. Integrating art into the curriculum had many benefits in understanding and comprehension in those content areas, but this method of art education did not allow for art to be seen as a serious discipline.

The standardized testing movement was further solidified in 2001 with the No Child Left Behind Act during the George W. Bush administration (Spring, 2005). Initially, this act created a nationalized school system and allowed variation in academic standards and content of state tests. However, the following years showed that these variations decreased. The National Assessment of Education Progress (NAEP) required that every other year a sample of fourth and eighth graders from each state would take the national standardized test. The National Assessment of Educational Progress would further have an effect on state testing by comparing the national test results with that of the state test results making state tests conform to the national standards of testing. From 2002 to 2008, the NAEP required more standardized testing in more content areas and more grade levels in all school districts. From 2002-2003, states were required to show annual report cards with student achievement scores and test scores by school district. Each school district was also required to show district wide scores and school to school scores. School districts were also required by the NEAP to give assessments biennial for fourth

and eighth graders in reading and mathematics. From 2005-2006, each state was required to have academic standards in reading, mathematics, and science for all public elementary and secondary schools. Each state was also required to give annual statewide tests in reading and mathematics for grade levels three to eight. From 2007-2008, states were required to give standardized science tests at least once in elementary, middle, and high school (Spring, 2005).

Not until recently had there been this scale of emphasis on art education. And even with this new found emphasis on art education and assessment, it was not national but rather certain states that have begun to take the next step in requiring an art education for all public school students and creating standardized assessments to evaluate those students and the art programs. In 1988, the California Department of Education decided that the state needed curriculum frameworks for content areas that were not previously assessed through standardized testing. California led the way in developing a history and social science framework including history, geography, economics, political science, anthropology, psychology, sociology, and the humanities (Herman & Winters, 1992).

Shortly after this, Washington State also determined a state framework in the form of Essential Academic Learning Requirements (EALRs) and Grade Level Expectations (GLEs). The Essential Academic Learning Requirements (EALRs) for the arts fall under four main categories with corresponding sub-categories. The four main categories are as follows: the student understands and applies arts knowledge and skills; the student demonstrates thinking skills using artistic processes; the student communicates through the arts; and the student makes connection within and across the arts to other disciplines, life, culture, and work (www.k12.wa.us).

In 2003, Washington State took another big step by incorporating the fine arts (visual

arts, music, theater, and dance) into standardized assessment. This took the form of the Classroom Based Performance Assessment (CBPA). The CBPA is "designed to assess the Washington state Essential Academic Learning Requirements...All items represent what we want all students to know and be able to do at the benchmark levels of grades 5, 8, and 10 (high school) in dance, music, theatre, and visual arts in alignment with our state arts EALRs" (www.k12.wa.us/assessent/WASL/arts/CBPA). The visual art CBPA is a criterion referenced assessment which means that teachers are looking for student understanding as opposed to comparing students to one another based on a moveable grading curve. Art EALR numbers three and four are integral to each CBPA even if those standards are not listed on the assessment. Classroom Based Performance Assessments are now used to ensure that students are getting key skills and knowledge in the areas of social studies, the arts, and health and fitness. These CBPAs are created by state teachers from different grade levels and different districts. The CBPA is also administered in the classroom by the teacher. Washington's goal is that "by the end of the 2008-2009 school year, school districts will have in place in elementary schools, middle schools, and high schools assessments or other strategies to assure that students have an opportunity to learn the essential academic learning requirements in social studies (includes history, geography, civics, economics, and social studies skills), the arts, and health and fitness...Beginning with the 2008-2009 school year, school districts will annually submit an implementation verification reports to the Office of Superintendent of Instruction"(www.k12.wa.us/assessment/WASL/arts/CBPA).

Only in recent years have educators begun to experiment and implement alternative, authentic assessment. There is little history and little research concerning alternative art assessment. The following chapter reviews research studies that examine multiple means of alternative assessment in different performance areas in schooling. Studies that question

portfolio assessment, self assessment and reflection, and peer assessment will be reviewed. Also noted are the validity and reliability of these forms of assessment. These research studies will provide insight into what the future could hold for alternative assessment in the visual arts.

Summary

The history of art education stretches from Greek antiquity to today. In the beginning, the arts were seen as an integral part of a fully rounded education that included reading, writing, exercise, and music. Over time, the opinion of art education has fluxed and changed nature many times.

In the Industrial Age, the arts were used to train individuals for the work force. Art was seen as a strong vocational skill. This resulted in an art education that was very rooted in business and education. As business and education continued to take hold in the nation's public school system, the role of art in the classroom changed. Art education was cast aside as the content areas of math and science took hold as the important education for a growing work force. This not only affected the importance of art education, but it also affected the funding of art education and the assessment of students as a whole.

The new focus on students as human capital created the opportunity for businesses and the government to define student success and how to evaluate that success. As a result, presidential administration passed several acts that required students to take standardized tests every year of their school career. In particular, Washington State created standardized testing for all content areas, including the visual arts. Washington State created several sets of standards of learning including Essential Academic Learning Requirements and Grade Level Expectations. These standards corresponded with required, state created performance assessments for all visual art students.

It has only been recently that educators and schools have been looking to alternative forms of assessment for the visual arts. The Washington CBPAs are a step in that direction. Even though it is a form of standardized assessment, it is a criterion referenced assessment that looks for individual student learning and understanding as opposed to grading a student based on his or her peers. Art educators and teachers are now looking at different means of assessment that can show student understanding in multiple ways.

The next chapter will examine research studies of different forms of alternative assessment in the visual arts.

CHAPTER THREE: CRITICAL REVIEW OF RELEVANT STUDIES

Introduction

Chapter One discussed assessment of the visual arts, and that assessment in the visual arts should be motivating, challenging, and engaging. It can be used as an integral part of a visual art curriculum that runs concurrent with student learning.

In the preceeding chapter, a survey of the history of art education and its accompanying assessment was shared that included opinions, ideas, and theories from ancient Greece, the nineteenth century, the Progressive Education movement, and the movement of educational reform to prepare students for entry into the competitive marketplace. All of these have shown to have had an influence in art education and assessment.

This chapter critically examines different research studies and literature on the development of alternative assessment in a visual art curriculum that is valid and reliable as well as the impact that this alternative assessment may have on student performance and engagement. The chapter is organized into four sections.

The first section involves portfolio assessment, the second section is concerned with self and peer assessment, and the third section is concerned with performance assessment methods, more specifically using rubrics and guidelines to assess a student's performance on a task. The final section of this chapter involves research that investigates the effects of assessment on student and teacher attitudes as well as the importance of curriculum development, teacher training, and teacher judgments while using alternative forms of assessment.

Portfolio Assessment

This section aims to address the effects of using a portfolio assessment as a means to assess student performance in a visual art classroom. The conditions that make portfolio assessment effective will be examined. Portfolio assessment can be defined as a purposeful collection of student work that tells the story of the student's effort, process, or achievement in a given area (Beattie, 1997). Beattie (1997) also elaborated on the different types of portfolio assessment that are available to art instructors for use. These include but are not limited to: a best works portfolio, an expanded art portfolio, a mini portfolio, and a process portfolio. For the purpose of this paper, the studies analyzed mainly employed a best works portfolio. The best works portfolio is defined as being inclusive of only a student's best work accumulated over a specific period of time. The portfolio in the art classroom is seen today as significant evidence of a student's progress, achievements, and experiences that encourages student-teacher collaboration, student research, and communication; confirms multiple learning styles and levels; promotes reflection and self assessment; and motivates students (Beattie, 1997). Portfolio assessment is used in more than just the visual arts; portfolios are now being used in other content areas as well including language arts, writing, and science.

Dorn (2003) conducted a quantitative experiment with 51 different schools in three different states to evaluate K-12 learning through art teacher assessments of student portfolios. More specifically, Dorn asked three questions: can you put a grade on an art portfolio, are teachers reliable or do they have the training to assess portfolios without bias and evenly across different classes, and how does this type of assessment relate to the tracking system rooted in a majority of school districts in the United States.

Seventy K-12 art teachers (nT=70) and 1000 (nS=1000) students from eleven different school districts in Florida, Indiana, and Illinois volunteered to participate in this study

over a four to eight month period. The study was divided into separate activities for the teachers: training in the use of art rubrics in assessing art performance, experiences in using blind scoring methods by peer teachers to validate teacher-scored student work, training in the use of authentically scored student art as a curriculum tool for the improvement of art instruction, developing assessment portfolios and analytical rubrics for special needs, and developing assessment instruments and methods of reporting consistent with student needs and with Goals 2000 and state school district standards. The design of the portfolio assessment study involved the use of repeated measures on the same multiple observations. The teacher in each school selected one class and performance assessment measures were applied on two portfolios from each student in each of the selected classes. The measures included three teacher ratings on each student art portfolio containing four art works gathered before and after the teacher training. Each teacher collected four student art works from the same class to form portfolio A, which was scored using rubrics on a scale of one to four by the teacher with two additional teachers blind scoring the same portfolio. These works were again scored along with four new works gathered at the completion of the teacher training by three teachers in the study group, forming portfolio B. Portfolio B was assessed by two different study groups labeled B1 and B2.

Teachers participated in eighteen different one and two day training sessions. The goals of these training sessions were to administer a field tested, authentic K-12 assessment model on student art work; to develop and test teacher designed assessment models for use in the participating school districts; to organize a system to collect data for the assessment; to report the data collected from the assessment in formats that satisfied individual school, school district, and state assessment standards; and to participate in selected studio experiences and curriculum development. The teachers were expected to create a process where teachers could

learn to accurately assess student art performances in the context of their own school district. They also participated in creating studio work that they would expect their students to know. The teachers' work during these training sessions was critiqued by the workshop instructors at the end of each session. Another critique was done with student work from the teachers' chosen classes. These critiques were led by the project director, the art supervisor, or the teachers themselves. Upon completing these sessions, teachers began constructing curriculum and lesson plans based on what they had learned in the training workshops.

The performance assessments were designed in such a way that they included: both the procedural and focal knowledge that students needed in order for them to know how and be able to complete various learning activities for the arts, the core performance roles or situations that all K-12 students should encounter and be expected to master, the most salient and insightful discriminators that could be used in judging artistic performances, sufficient depth and breadth to allow valid generalizations about student competence, the training necessary for those evaluating artistic performances to arrive as a valid and reliable assessment, and a description of audiences that should be targeted for assessment information and how that assessment should be designed, conducted, and reported to those audiences. The most important concern of the assessment model was that it reflected the nature of the exercises already embedded in the art curriculum and that it encouraged students to study their own train of thinking as revealed in notes, sketches, or artistic practice.

A holistic rubric was also completed with the same detail. The rubrics were used to assess different performance levels: excellent, very good, satisfactory, and inadequate. The rubric descriptors for each grade level reflected the age appropriate cognitive, aesthetic, and technical skills needed. The specific performances described in the rubric came from three different theorists and their corresponding ideas on student development. The first source was

Piaget's pre-operational, early concrete operational, and formal operational stages. The second source was Lowenfeld's different stages of scribbling, pre-schematic, schematic, gang stage, reasoning stage, and period of decision stage. The last source was McFee's skill improvement states, which includes: searching for pattern; using verbal descriptors of space; exploring consistencies in shape, form, and size; manipulating objects as a unit; taking an average; completing visual wholes; and recognizing patterns in figure and ground.

For inter-reliability, teachers were advised that a difference of one point was acceptable. A score difference of two or more points suggested that perhaps judges were not viewing the same features in the student artworks. Teachers were also warned of reader fatigue - when judges became unfocused because they were too tired or needed a break. When this appeared to happen often, the judging was stopped and the teachers, either as a group of individually, took breaks. Inter-reliability revealed all the correlation coefficients were medium to low, but all were significant (A1/B1=0.552, A1/B2=0.345, and B1/B2=0.442). In determining whether the scores for any classroom were distributed evenly and whether there was a sufficient score spread to determine whether the test discriminated among the portfolio scores, the results revealed that the score spread for A1 was smaller that the score spreads for B1 and B2. When comparing the scores in the B2 pre-test with the scores in A1 and B2, the B2 pre-test had a greater range and a more even distribution that A1 or B2. This suggests that teachers, acting as evaluators, were more discriminating when comparing the B1 and B2 portfolios (Dorn, 2003).

In the area of student performance, sixteen schools showed an increase in the class mean score by more than 50%, eleven schools showed a decrease in the class mean scores. These declines ranged from as low as 0.03 (3%) to 0.50 (50%). Dorn (2003) noted that the decline in the mean score for the eleven schools does not necessarily mean that student

performance decreased from the beginning of the experiment. More than likely, lower student scores occurred when compared to many students that showed improvement. The results also showed that the greatest increase in student scores happened at the 6-8 grade levels, somewhat in the 1-3 grade levels, and much less in the 4-5 and 9-12 grade levels. Even more interesting was the difference in the improvement of students of different academic levels. While less than half of the higher achieving students improved their scores, over 85% of low achieving students improved their scores.

While these results support the idea that portfolio assessment can be a reliable means of assessing students, the participants were all volunteers in the study. Each teacher was not randomly selected and each teacher came from 51 different schools in 15 different school districts, each with their own context, populations, resources, and school support levels. While this does add to the generalizability of the study, there is also the importance for teacher training. While all these teachers came from differing backgrounds, they were all subject to the same training and support from the researcher and the researcher's team. All these teachers were familiar with the nature of art and evaluating art with similar criteria, even if state/local standards were different. This note demonstrates the importance of teacher training, attitudes, and scorer reliability which will be discussed further in this chapter.

Sandi Uram (1995) questioned if sixth grade students had enough knowledge to evaluate their own artwork by using portfolios. Uram targeted twelve sixth grade students (n=12) in an elementary magnet school. The school's curriculum was developed by the district and art specialists were expected to use this curriculum as a guide when developing their own lesson plans. The art specialist was a 3/5 teacher that taught art classes as a place for students to be while other classroom teachers have a planning period. This particular art specialist taught six classes a day. This particular school district wished to address the issue of arts assessment by

reforming the curriculum. Uram (1995) observed the implementation of an alternative assessment plan from August 1994 to December 1994. The assessment plan consisted of four different categories: revise the curriculum, introduce portfolio assessment into the current curriculum, increase student involvement by using self and peer assessment, and improve communication between the teacher and student. Three methods were used to assess the effectiveness of the portfolios. Teacher evaluations, as well as peer assessment and self assessment forms reflected the expectations of each art activity. The portfolio assessment was intended to reveal the growth of the student as documented by completed art works. A post survey was used to measure students' ability to assess their own growth.

Prior to this new plan of curriculum, a pre-survey was given to both teachers and students. The students expressed low confidence in their ability to assess and evaluate their work; and often, students were unable to see and document their growth in art. A survey was given to eighteen classroom teachers. Only five returned the survey: one teacher discussed her love of weaving projects, one drew a frown, and the other three said nothing. At the beginning of the school year, teachers focused on cooperative learning to teach social skills and group bonding. This was done to build a strong foundation for using portfolios in the classroom. As the class progressed, students completed four different projects in regard to their portfolios: creating a folder to be their physical portfolio, creating an art book while working in groups, using recycled plastic containers to create African Boli, and creating pieces of jewelry out of small, lightweight materials. At the end of December, the teacher held individual conferences with each student. Together they noted the student's strengths and weaknesses and set goals for improvement. During these conferences, a majority of the students expressed a desire to continue using portfolios and showed more pride and ownership in their work.

A post survey was given to students at the end of the experiment. The survey results

revealed that more students felt an increased ability to relate what they learned in art to other content areas and students' skill development improved through activities and lessons in meta-cognition, cross curriculum learning, multiple intellligences and cooperative learning. Students also expressed enjoyment sharing ideas with classmates. The survey results showed that student work improved because of the combination of cooperative learning with meta-cognitive behavior (Uram, 1995).

The very small sample number makes it hard to draw any solid conclusions from this study. There was also no mention of how the students in the study were actually graded. Teacher opinions from the pre-survey were very grim as well. Their responses sent up a warning flare that the teachers who actually did participate in the study may not have given their full focus or attention. The researcher also stated that the participants did not have any curriculum in place that addressed assessment, and the educators failed to understand or appreciate the value of using art for a student's academic growth and achievement.

Portfolio assessment was also used to document the authentic problem students created as a graduation project for a mechatronics class in a high school in Israel. Doppelt (2009) investigated the role of portfolios as a means of showing student progress in their graduation projects. One hundred and twenty-eight high school students (n=28) in Israel who have studied mechatronics from grades 10 to 12 participated in this case study conducted over the course of seven years. There were three stages of the experiment: 1) field research was designed to begin the creative design process that helped students in designing creative and authentic projects, 2) understand the way students designed their projects, and 3) field test the creative thinking scale as a means to assess creative design process and validate it with the results of external assessment. The students had to complete a graduation project in the 12th grade. Data was collected through natural observation and recorded in the researcher diary (the researcher was

the tutor that was with the students from 1996 to 2003). The researcher documented the students' progress, problems, and evaluative criteria developed in the classes. Students finished with an exam in the form of a portfolio assessment by an outside instructor and a 20-minute presentation in front of the instructor. The researcher found that students thrived on authentic projects, one that is a student's own idea chosen out of curiosity and genuine interest. The students were devoted to their project in all of their free time to research any necessary information as it was not provided by the class curriculum. Educators also need to identify skills the students gained through the creation of their projects. It is also important to identify what skills the students are less competent with. This could be done using the creative thinking scale. In the examination, all the students received a passing grade above 80 (the passing score was 55). Most students also created portfolios that reflected a high level of achievement. Through creative process and reflections, students saw their own progress and directed their thinking as needed. The projects showed that students created their own authentic assignments and learned from them.

No real conclusions can be drawn from this study however. While the researcher was with the subjects for an extended period of observation, the researcher failed to include any history of the study that could have affected the results. The researcher failed to address any bias that could have occurred during his time with the students. Doppelt (2009) also failed to elaborate on the observations and what those observations meant to the study other than one brief mention of a group of students that created an authentic project. This section of the study was very brief and did not get into detail on the students' thinking process. The section described what they did, the problems that arose, and how they solved them with no description of student thinking of the steps between. The researcher failed to give more information on his own subjectivity and personal place in the experiment. The researcher in this

case was also the tutor for the students. There was no mention of the criteria used to evaluate the students, how the students were evaluated, and what questions were asked of this student. This study left many unanswered questions with very few determined conclusions.

Shober (1996) conducted a quantitative experiment in which twenty two students (n=22) created a portfolio to demonstrate student growth over the course of twelve weeks. This portfolio was used as an evaluation tool for parent/teacher/student conferences. Shober's two main goals for this experiment were that 50% of the students would show improvement in narrative writing by evaluating their portfolio samples on a rubric scale and 60% of the students would participate in a parent/teacher/student conference with their portfolios as a way to show student progress. Over the twelve weeks, the researcher worked carefully and closely with students to build a safe community and guide the students in the writing process to complete three different narrative writing pieces for their portfolio. These pieces were assessed for growth and understanding of the writing process. The researcher was also the classroom teacher. The students were able to feel relaxed and confident in the process and with the teacher and each other. This allowed for more openness creating their portfolio pieces. However, there was little teacher observation of the student's cognitive growth or motivation during class time. There was also little mention of the teacher's specific means of teaching students how to write for their portfolios. This study did demonstrate the importance of teacher training to evaluate portfolios, but there is no evidence for whether portfolio assessment is seen as valid or reliable.

Teachers and parents were given questionnaires to determine their opinions and ideas on portfolio assessment. Both sets of surveys consisted of ten questions with five different choices: always, usually, sometimes, infrequent, and never. The parent survey questions were as follows: 1) have you been involved in a conference in which a portfolio was utilized to show your

child's progress, 2) do you feel that portfolio assessment could be valuable in a parent/teacher conference, 3) do you spend a brief time asking your child what transpired in school on a daily basis, 4) do you feel that portfolio assessment would adequately present your child's progress, 5) do you feel you experience good communication between parent and teacher during your conference, 6) do you feel adequate time is provided by the teacher to discuss your child's progress and your concerns, 7) would you feel comfortable attending a student/parent/teacher conference, 8) would you like to see writing samples included in your child's portfolio assessment, 9)would you be willing to have input into your child's portfolio, and 10) do you believe that you, as a parent, are empowered to help your child succeed in school.

Parent answers were figured into percentages for each possible answer. The researcher focused on the highest percentages for her experiment. Thirty percent (30%) of parents stated that they had never participated in a conference that utilized portfolio assessment while another thirty percent stated that they infrequently experienced portfolio assessment during conferences. Seventy-eight percent (78%) of parents felt that portfolio assessment could always be valuable in a parent/teacher conference. Fifty-seven percent (57%) of parents stated that they always asked their child about school on a daily basis while thirty-five (35%) percent stated that they usually asked. Forty-eight percent (48%) of parents felt portfolios could always adequately present student progress while thirty-nine (39%) percent stated usually. Fifty-three percent (53%) of parents stated they usually experience good communication at conferences. Thirty-nine percent (39%) of parents felt that only sometimes was adequate time provided to discuss their child and their concerns. Seventy percent (70%) of parents felt very comfortable with attending a student/parent/teacher conference. Seventy-seven percent (77%) of parents always wanted to see sample of their child's writing in a portfolio. Seventy-one percent (71%) of parents were always willing to be actively involved and provide input in their child's portfolio.

Ninety percent (90%) of parents felt that they were always empowered to help their child succeed in school.

Teachers also completed a survey consisting of ten questions. The questions as well as the teacher answers are as follows: 1) could assessment be defined as the process of gathering evidence and documenting a child's learning and growth with 54% always and 32% usually, 2) are portfolios workable tools for assessment with 32% usually and 43% sometimes, 3) do you feel portfolio assessment can be unique for each student with 32% always, 32% usually, and 26% sometimes, 4) do you believe that portfolio assessment can emphasize what a student knows with 21% always, 43% usually, and 32% sometimes, 5) do you think portfolio assessment can be utilized to present different developmental levels with 46% usually and 29% sometimes, 6) do you believe that portfolio assessment will enhance parent/teacher conferences 43% always and 32% usually, 7) do you feel portfolio assessment shows progress through product samples 36% always and 36% usually, 8) do you feel portfolios should be passed through grade levels with 43% always and 21% usually, 9) would you use the portfolio assessment as a means of communication in a parent conference with 49% always and 32% usually, and 10) will you take the time to compile and use a portfolio effectively with 38% always, 25% usually, and 29% sometimes.

Based on the answers from both parents and teachers, both parties felt that portfolio assessment was a beneficial component of a parent/teacher conference. They also felt that portfolios could be a tool for assessment that is unique to each student, shows student progress and growth, and involves parents in their child's learning. Overall, the students writing skills improved; 68% of students elevated their scores, 27% of students remained roughly the same, and one student decreased in scoring. Fifty-five percent of the students participated in a parent/teacher/student conference with their portfolios.

This study has many areas to note. Shober (1996) was able to create a safe working environment where students could feel relaxed and confident in this new assessment process by being both the teacher and the researcher. The students were able to feel more at ease with the teacher and with each other. Shober (1996) greatly detailed the background of the students, from ethnic to socioeconomic status to academic achievement. This is another example, however, of a study with a small sample number. Shober (1996) does not mention the rubric that was used or how that rubric was created. While the researcher was able to create a safe community by also being the teacher, there was also the question of researcher bias involved in the observations. At times, it seems as though Shober(1996) was reaching for conclusions that she was hoping to find as opposed to explaining her observations. There were few observations that mentioned student growth, progress, or attitudes in this study. Shober (1996) saw portfolios as a great way to monitor student progress and take a student from one grade level to the next as a means of formative assessment. While this sounds like a great idea, there was no real data to support whether or not it was effective. Reviewing the survey questions and results, there were few things that can be inferred, many of them were judgments and opinions. These do not reflect any answer for whether portfolio assessment is valid, reliable, or effective.

Shober (1996) did raise new questions concerning parent involvement in their child's assessment and academic growth such as: what were the parents' thought processes concerning portfolio usage, were some parents uncertain that portfolios were adequate as an assessment, and how did parents that participated in the conference actually perceive the validity of utilizing a portfolio approach to present the student's progress.

A quantitative study focused on what one hundred and seven students (n=107) from Canada (n=61), England (n=17), and the Netherlands (n=29) valued about portfolio assessment as well as their perceptions of portfolio assessment as a valid preparation for their future

(Blaikie, Schonau &Steers, 2004). The aim of the study was to compare Canadian, English, and Dutch student's opinions and experiences of portfolio preparation in art and design for final assessment in the terminal year of secondary school. Students participated in creating a portfolio based on one theme for the school year. Teachers were trained to assess these portfolios with similar rubrics and expectations. Researchers collected data on student opinions about assessment and students' actual experiences of portfolio assessment with survey-questionnaires based on a five point Likert scale, ranging from "strongly agree" to "strongly disagree" and "highly accurate" to "highly inaccurate," respectively. Researchers designed questions to distinguish between what students thought best practices should be and what they had actually experienced in participating in a portfolio assessment. The questions were separated into two different surveys: if it is important to understand criteria for assessment and if the participants themselves understood the criteria used to assess them.

Students agreed on a few things: the benefit of group critiques (Canada had 48.3% agree, England had 58.8% strongly agree, the Netherlands had 48.3 agree), more than one person assessing their work to combat bias (Canada had 54.1% strongly agree, England had 82.4% strongly agree, the Netherlands has 69.0% strongly agree), it is important for teachers to hear students' opinions on their own work (Canada had 86.9% strongly agree/agree, England had 100.0% strongly agree/agree, the Netherlands had 93.1% strongly agree/agree), importance of peer discussion (Canada had 68.8% strongly agree/agree, England had 88.3% strongly agree/agree, the Netherlands had 82.8% strongly agree/agree), preparing a portfolio is a useful learning experience (Canada had 90.1% strongly agree, England had 93.1% strongly agree, the Netherlands had 88.3% strongly agree), preparing a portfolio is a worthwhile experience (Canada had 91.8% strongly agree, England had 100.0% strongly agree, the Netherlands had 79.3% strongly agree), and high school art and design should be a good foundation for future

study (Canada had 86.7% strongly agree, England had 88.3% strongly agree, the Netherlands had 62.1% strongly agree).

The students' experiences of using portfolio assessment in their high school art and design classes differed however. English and Dutch students disagreed with Canadian students on the topic of being able to give input on assessment criteria with 34.4% of Canadian students stating they were able to input criteria as opposed to 58.8% of English students and 51.7% of Dutch students having no input on assessment criteria. The researchers believed this could be due to the fact that assessment in these countries is determined by national government agencies. Canadian students believed that luck played a major part in art assessment with 13.1% highly accurate and 19.7% accurate. Also, Canadian students saw this assessment as less reliable with only 44.3% of Canadian students agreeing that portfolios were useful as compared to 76.5% of English students. The researchers speculate that perhaps these two findings are due to limited experience with group critiques and clearly defined assessment criteria. However, the Canadian students were still fairly positive about portfolio preparation as a useful and worthwhile experience and continuing to study art and design further in the future when asked their opinion.

The students all agreed to a great percentage that this portfolio exercise helped them with their art works. Canadian students stated that portfolio assessment helped them with 37.7% answering "highly accurate" and 44.3% "accurate." English Students answered 58.8% "highly accurate" and 23.5% "accurate." Dutch students answered 17.2% "highly accurate" and 34.5% "accurate." The students were more engaged and more critical with their art work by sticking with a single theme and exploring that theme throughout the school year. However, the low sample size makes it hard to generalize these findings. The researchers did not give a complex view of the students' experiences, thought processes, interactions, or portfolio art

works. The researchers also did not present in full detail the rubrics used to assess the students.

There were some outliers in Blaike, Schonau, and Steers' s (2004) study. Blaikie returned to this study with curiosity concerning one student. In a qualitative case study, Blaikie (2008) conducted interviews with one student whose experiences were so vastly different from the other students. The central question was "what was your lived experience of being a high school art student?" Blaikie (2008) was looking to explain why it was that this one student's classroom and assessment experience was so different from the other participants. The researcher was also investigating what information was presented to the student as art knowledge and the student's feelings on this; in essence, questions concerning curriculum, assessment, and teaching methods. It seemed that this student's instructor was the worst art teacher in existence…or very close. According to the student, "Mr. Robson did not prepare formal art lessons, and consequently had no concrete expectations of his students with regard to learning in art"; "…there was no consequence for not doing any work stating that assessment was based on the instructor's subjective opinion"; "Art should be like math…one should assess based on concrete evidence of products and process, not perceptions of effort and likeability"; and "there were no explorations or discussions on cultures, art styles, history, or social issues." This was in contrast to what universities were looking for in portfolios. Universities were expecting a range of subjects, styles, media, and experimentation. The students stated that these had to be learned independently and the portfolio was created by using art works completed in other classes, at other times, and at other art workshops and events.

So what does this mean then for the visual art curriculum and assessment? Even if a student answers that portfolios are a great means of preparation, self reflection, and presentation; it will not be worth anything if the assessment and the curriculum do not match. There needs to be teacher support, scaffolding, and relevant classroom content for students to

feel that their portfolios are worthwhile and meaningful.

In a study on student motivation and learning with students (n=51) from five different schools in Portugal that began using portfolios as a means of final examination (Pereira, 2005), researchers noticed the importance of the teacher to the process of creating and compiling a portfolio. Portfolios were used as the new examination for secondary art students (US equivalent). The students were allowed to choose a theme and then explore, plan, complete, present, and evaluate their work. These students and their teachers (n=7) were then interviewed and given surveys and questionnaires to determine their viewpoints on student engagement, motivation, and critical thinking. Students considered the new exam method valid and enjoyed the freedom it gave them. Students stated, "the portfolio was good essentially because we had to make different works; it was not prescriptive; we had a theme and we had to develop the work...from our heads; not like go and draw, draw a bench with a monkey...I did a portfolio with things I like to do and showing what I wanted to know." The portfolio was also a way for students to demonstrate their creativity. One student claimed, "the portfolio demands more creativity; it is much more enjoyable and motivating for us; but it is also more difficult because of that." Students also saw this form of assessment as authentic. "The portfolio was more real than the usual work and tests we do in the art classes; it is about our life, not school life or tasks that school thinks are important for us."

Student saw portfolios as an authentic means of assessment. They were motivated, excited, and engaged. However, this was dependent on the teacher's view of the students, the assessment, teaching methods and prior knowledge and experience using a similar assessment method and though process in the arts. Different teachers reported different findings, respectively: "They have a passive attitude towards art disciplines because they are used and were trained as passive objects to make pre-determined tasks"; "The main difficulties in

implementing the portfolio with my art class were related to the rationale of the portfolios. It was very difficult to require students to think independently because they are used to following detailed prescriptions for each task”; "Since we started the portfolio in September, students had the time to be prepared. Little by little, they learned skills of critical reflection and self-evaluation. They learned how to plan the different steps of the work in the time to respond to the deadlines. It was good to continue to use portfolios after the first experience at the beginning of the year. Students acquired so many learning experiences that allowed them the opportunity to develop organized ways of thinking and making. They developed independent skills and little by little they pushed their own barriers. They learned how to realize their own intentions and how to make projects that motivated them." It should be noted that this school was the pilot school for this experiment. The students had more experience working with the portfolio assessment than the four other schools which probably resulted in the students’ increased preparation and learning.

We can see a definite contrast between the interpretations of teacher's thoughts on portfolio assessment in relation to the scaffolding, teaching methods, and teacher attitudes. One teacher even stated that her students simply did not have the capacity for this style of thinking and assessment. Students that were allowed time to build the skills needed to participate in this kind of assessment excelled and were able to organize their thinking, experiment with art processes, and challenge themselves.

Pereira (2005) selected participants from varying geographical locations and resource levels. This study also took place over an entire school year so there was plenty of time for students to become accustomed to the assessment method. The researcher collected the data in a very clear manner and the criteria for evidence needed in the portfolios were standardized for all the schools. The criteria was formatted in a very easy to read chart. The criteria evidence consisted

of preliminary studies such as sketches and journal notes; investigation material including both written and visual critical inquiry; self assessment reports in the form of interviews; records about the student's intentions progress reports; presentations; evaluation reports and critiques; reports of notes about previous experiences and interests both written and visual; and final visual products such as paintings, drawings, sculptures, graphic design, films, videos, and installations. This portfolio was a very inclusive, well rounded example of student progress and interest. The study samples were volunteers which could have had an effect on the study. By using volunteers, the researcher can assume that fewer or no participants will drop out of the study, affecting the history. This does effect the history of the study, but it also effects the opinions, backgrounds, and choices of the teachers. These participants more than likely had a greater interest in art and portfolio assessment in contrast to randomly chosen participants that could have varied in interest level and experience. Different activities and tasks during the class time could have also affected the students' perceptions about portfolio assessment. The teachers' attitudes also affected how student felt towards their achievement and the portfolio assessment. Pereira (2005) did not point these out in full detail. However, she did make the note of one teacher's negative attitudes toward the students as having a negative effect on the student's perceptions of themselves as learners. Once again, the theme of the teacher as the driving and deciding force of assessment was seen.

Summary

This section analyzed the validity and reliability of a portfolio assessment method. Given the findings, it can be suggested that portfolio is an assessment method that is both valid and reliable in the correct curriculum with the correct instruction. Students and teachers both felt that portfolio assessment could increase student motivation, growth, and self awareness through the process of collecting, analyzing and reflecting on their thought process and their final product. Over time, students and teachers began to feel more comfortable with this assessment method.

The next section reviews the effectiveness of student self assessment and reflection as well as peer assessment. The two main areas of student self and peer assessment that are analyzed here are the use of rubrics and involving students in their own learning through records and student created assessment.

Self-Assessment

According to different motivation theorists, self assessment can contribute many positive feelings about a student's own learning, choice, and self worth (Brookhart, 2004). Self assessment can be defined as students taking responsibility for monitoring and making judgments concerning their own learning (Somervell, 1993). Somervell (1993) also stated that self assessment encourages students to look to themselves and to other sources to determine what criteria should be used in judging their work instead of relying solely on their teachers. Research has shown that students are highly motivated when they observe the progress they are making and strive to reach their goals.

Peer assessment can be defined as an assessment carried out by fellow students. It can also be part of the self assessment process and serves to inform self assessment. Other

students' input can be very useful in the assessment process. Peers can observe one another in their learning process, and often times have more understanding of their peers' work than the teachers.

Even more so, self, peer, and collaborative assessment are seen as a student centered, democratic classroom practice that removes the barriers between student and teacher and builds greater confidence in students that can lead to motivation and deeper learning (Somervell, 1993). This section will investigate studies that concern self and peer assessment.

Brookhart (2005) questioned if student self assessment and self recording could improve a math memory task into a learning exercise that could contribute to a student's mathematical literacy development. This small study of two different third grade math classrooms included 21 students (n1=21) in one class and twenty students (n2=21) in another. The project involved three student teachers, two university supervisors and three cooperating teachers with two classes of students at a suburban elementary school in the Eastern US. Two of the teachers were regular third grade teachers and the third was a special education teacher that worked in those classes. Students included both general education students and students with Individual Education Plans. The school at the time of the study had an enrollment of 425 of which 1.6% were classified as low-income students. The school also consisted of 24 professional staff.

Five- minute timed multiplication fact tests were given once a week for ten weeks. The students practiced with times tables and practice with strategies such as flash cards, games, and student partnering. Students also completed a prediction exercise and a reflection sheet with each test. Students graphed their actual score next to their predicted score and predicted their score for the next week. They used a reflection sheet to write whether they had met their goal from the week earlier, what strategy they used and how well it worked, and what strategy or

strategies they planned to use the next week.

On average, both classes predicted their achievement very well. The overall average for both classes progressed as the weeks went by. Class one reflections revealed that flash cards and timed tests were the most commonly used strategies. Practice and memory were the attributions students most often suggested as the reasons for success. Lack of practice was the reason most students stated for not feeling their strategy worked. Class two stated that repeated addition, flash cards, writing sentences and studying with parents were the most used strategies. About half the students expressed the intention to learn the math facts rather than to get high test scores. Teachers said that they noticed their students' primary learning came from seeing how they did on the test, receiving feedback in a consistent manner, and applying what they knew about bar graphs to real life. They thought they had learned more about bar graphing as the students did not need instruction to know where to put their dots for their grades and predictions. Some students had difficulty with the reflection writing sheets. Teachers stated that they needed to teach them to be confident in their own thought processes. They also stated that they thought their students learned what worked best for them in studying strategies. One teacher stated that she believed that students learned the multiplication tables better the year of the study compared to the year before.

No sure conclusions can be drawn from this study though. This study too had a small sample size and was only comparing two classrooms in the same school at the same grade level. This means that no general statements can be made about self assessment from this study. Another major weakness in the study was the history involved in the classrooms because one class's students were having difficulty understanding the reflection worksheet, the teacher changed the worksheet to fit her students. The researcher did not go into detail about the changes to the worksheet or how the original worksheet compared with this revised version.

Also, concerning the reflection worksheet, some students admitted to simply filling in the blanks and not fully understanding how to complete a reflection. The students should have been given standard instruction on how to complete these reflections.

Brantmeier (2006) conducted a quantitative study to determine if a self assessed rating could adequately predict student achievement and reading performance. This was accomplished through a pre-test/post-test method. The participants present for this study were seventy one (n=71) students ages 19-22 enrolled in an advanced level Spanish grammar and composition course (Spanish 301) at University X. This Spanish course was a third year course taught by five different instructors. It was also the first course of a two part sequence that was to be taken immediately before entering the Spanish literature courses at the university. Students were assigned to read long literary works for this course. Before taking this Spanish course, students were to take the Romance Languages and Literatures Online Placement Exam (RLL OPLE). This placement test was first developed and used in the summer of 2001 and had been administered on a regular basis ever since. The majority of students that take this exam were entering freshmen. Immediately after the exam, results were reported to the individual student and to the student's instructor to determine the student's language level.

To ensure a homogenous population, only students that satisfied the following criteria were included in the analysis: students who had achieved the appropriate composite score on the OPLE, students whose native language was English, students who enrolled in Spanish 307 the semester immediately following the OPLE, and students who completed all tasks for all data collection settings. In the end, this population totaled thirty-four students (nfinal=34). At this particular university, language courses were not required; therefore, the participants had signed up for this class voluntarily.

The participants answered questions before taking the OPLE and after in class readings.

A five point Likert scale was used to answer two pre-test questions: how well can you read in Spanish and how do you rate yourself as a reader of Spanish. This pre-test was used as a general assessment. The readers of these questions completed them online before beginning the timed placement exam. Two post-test questions were also created for students to answer after their reading they had to read for class. The first question was answered using a statement scale from "I strongly disagree with this statement" to "I strongly agree with this statement." That first question was "I found the passage I just read easy to understand." The second question was concerned with how much the students understood of the passage they had read. The answers for this questions ranged from "I did not understand very much at all" to "I understood all of the passage."

The reading section of the online placement exam was one variable. These reading included excerpts and writings about the daily lives of students, historical vignettes, poems, narratives, and encyclopedia type reading. The reading section for the classroom was carefully chosen after examining the reading materials appropriate for this level of language education. Before this study, the reading passage chosen was piloted with 67 students to ensure that the students were all familiar with the topic of the reading. All the assessments in this study were completed in the participant's native language. During a regular class period, the students completed the following tasks in this order: reading passage, self assessment questionnaire, written recall, sentence completion items, multiple choice questions, and a topic familiarity questionnaire. Both the researchers and the course instructors were present during the time the data was collected to ensure that the participants could not look back to read previous passages while completing the tasks.

The independent variable was the self assessment rating. For the reading section of the OPLE, researchers used the total number of correct responses as the dependent variable. The

reading performances that followed consisted of three different dependent variables – recall with an inter-reliability index of 0.96, sentence completion with an inter-reliability of 0.94, and multiple choice for which there was only one correct answer for each question.

The results of this study revealed no significant associations or correlations between a reader's self assessment and OPLE score; nor any significant correlations between the reader's self assessment and the reading performance as evaluated through recall, sentence completion tasks, and multiple choice questions. There were also no associations between post self assessment and the performance tasks of recall, sentence completion, and multiple choice questions. In examining the mean rating between the pre-test questions and the post-test questions, there was no significant difference as most students pre-assessed themselves as "ok" readers of Spanish, and then self-assessed themselves as being able to moderately understand a passage of reading. No conclusions can be drawn from this study.

The research did suggest areas for future study as this study revealed inconsistent findings. More research on self assessment as an indicator or ability and academic achievement could be conducted. Brantmeier (2005) suggested a future study to examine instructional practices that improve self assessment and also whether those practices improve student performance. Brantmeier (2005) speculated that students that are more prepared to evaluate themselves through in-class practice may achieve more in the classroom. This study was not an attempt to examine a student's metacognitive skills for reading. Likewise, those assumptions should not be made based upon this study. This study acted as an initial test on the reliability of self assessment to predict a student's future placement and performance. The findings showed that an early self assessment as a prediction is not a dependable measure of a student's ability in a classroom.

Hassmen and Hunt (1994) conducted an experiment to determine if a self assessment method with a multiple choice test would more accurately measure the test taker's knowledge. Hassmen and Hunt (1994) stated several reasons for using this self assessment method in a multiple choice setting: 1) to make a multiple choice test more accurate and more inclusive of a test taker's knowledge, 2) give more credit to the test taker that is sure of his or her knowledge, and 3) allow the test taker to express doubt or certainty about their answers. These three different reasons could shed some light on the effects of multiple choice tests concerning cultural differences, gender bias, and test anxiety. Hassmen and Hunt (1994) were interested in whether making self assessments regarding the correctness of an answer affected the total number of correct answers.

Of the 120 undergraduate students (n=120) who signed up to participate in this study, 60 were male and 60 were female. These participants were randomly divided into sub-groups, each group having either 30 males or 30 females.

A fifty question multiple choice test was created. The questions were chosen from old SAT test questions and were intended to be as close to gender equal as possible. A piloted version of the test revealed that math questions took up considerably more time than verbal questions. As a result, ten questions tested mathematic ability while forty questions tested verbal ability. Each question had five possible choices with only one being the correct answer.

The first two subgroups, 30 male and 30 female, completed the multiple choice test by answering the questions on a conventional multiple choice answer sheet. The remaining two subgroups (30 male and 30 female) answered the same test questions on a similar choice of self assessment answer sheet. These participants were asked to assess how sure they were of their answers. This self assessment was done immediately after each answer on the same answer sheet by marking on a five point scale (e.g. Almost a Guess, Probably a Guess, Neutral, Fairly Certain, and Almost Certain). The number of points gained or lost depended on whether or not the answer was correct or incorrect. This number was logarithmically related to the level of sureness. As a result, the participant who was fully confident and fully informed got a 100% correct answer and a 100% on the self assessment score.

Forty students were given the test at a time, ten participants from each of the subgroups. Participants received standardized written and verbal instructions before the test. The verbal instructions were the same while the written instructions differed slightly depending on whether the participants would self assess their answers or not. The dependent measures were the number of correct answers and the self assessment scores.

The performance of females, in regard to self assessment, was immediately affected by asking females to assess the correctness of their test responses. However, the results also seemed to indicate that only females benefited from self assessment. According to Hassmen and Hunt (1994), the subgroup of females that completed the multiple choice test only differed

significantly compared to the subgroup that participated in the self assessment ($p<0.01$). The mean total for the Female MC subgroup was 23.74; whereas, the mean total for the Female SA subgroup was 27.74. Concerning the male participants, the mean total for the Male MC was 29.19 and the mean total for the Male SA was 29.90.

The researchers developed one hypothesis to explain why females were affected but not males. Hassmen and Hunt (1994) speculated that these different results could be due to different cognitive styles of the genders. If females relied more on recall, asking them to assess their answers could have caused them to more carefully analyze their choices. On the other hand, if males relied more on problem solving than recall, they might not have been affected by the self assessment. They would have already been confident in their answer due to actively working to problem solve.

Additionally, male and female self assessment scores were not significantly different. The Male SA scored a 74.0 as compared to the Female SA score of 73.1. If one group had been more accurate in their self assessment, this would have been observed in the Percent Self Assessment scores. Males did score higher on the Sure and Correct score with a mean of 30.7 compared to the female score of 22.9. This suggests that males were better able to identify a correct answer than females once that answer had been selected. Females did have a higher percentage in the Unsure and Wrong category, indicating that they were more uninformed of a topic on the test than males. By using the extra information from the self assessment, Hassmen and Hunt (1994) were able to distinguish between participants that were misinformed (high percentage Sure but Wrong) and participants that were uninformed (high percentage Unsure and Wrong). This speaks to self assessment as a means to inform teachers and instructors of information that students are not comprehending and why they may not comprehend that information.

This study raised interesting questions regarding the nature of gender is self assessment. Females were shown to receive more benefit from a self assessment with a multiple choice test than male students. This phenomenon could have been a result of the difference between female learning patterns and male learning patterns, yet this study is inconclusive of this hypothesis. In regard to art assessment, it raised questions concerning both gender bias in visual art assessment methods as well as classroom instruction practices. As with any assessment method that is used in the classroom, the instructor must notice if there are any biases present, including gender bias. The visual art classroom can present a problem when dealing with gender bias. Two of the most common problems are the artists students study and the stigma surrounding art as a way of thinking. A majority of noteworthy artists throughout history are male. And with that, the visual arts is seen as a more feminine field when looking at craft or hobbies. However, the major industries that use visual art such as engineering and architecture, are male dominated fields. Hassmen and Hunt (1994) brought us a very good point: males and females might have different learning styles and these different learning styles should be taken into account when planning assessment.

While this study did randomly group its volunteered participants, there were not enough participants for generalizable conclusions. Along with this, the results did not allow for any sure conclusions in the question of assessment methods in the visual arts. This study did continue to demonstrate the theme that instruction methods and assessment methods should be carefully analyzed and scrutinized to prove if those methods are appropriate for a specific classroom.

Schunk (1996) conducted two quantitative studies to investigate how goals and self assessment affected student motivation and achievement outcomes. Schunk (1996) referred to self assessment as a three stage process consisting of self observation, self judgment, and self reaction.

Schunk's two studies (1996) were conducted in sequence. These studies investigated the self regulatory process among students during cognitive skill learning. The first study investigated the effects of giving students goals for learning or performance outcomes and examined the self evaluative process. Goals can be used as standards for students to see growth in their performance. Schunk (1996) also believed that setting goals would provide students with a sense of ownership and motivation to obtain those goals. This can also increase student engagement in classroom activities and tasks.

The final sample of participants consisted of forty-four fourth grade students (n=44) from two classes in one elementary school. The children were predominantly middle class with 18 girls and 26 boys, 24 White and 20 African American. Students were in regular mathematics classes and the administration described the students to be academically average.

The study consisted of both a pre-test and a post-test. The pre-test consisted of measures of goals orientation, self efficacy skill, and persistence. This pre-test was administered by an outside tester.

Schunk (1996) defined goal orientation as a group of intentions that affect how students engage in learning activities. These goal orientations were examined to determine if the goal and the self evaluation practice created different effects on a student's preference to different classroom goals. The orientation included 18 different items that tapped into four different goal orientations: task, ego, affiliative, and work avoidant. The student decided how well these 18 items described how they felt during a math activity. Students used a 10 point scale ranging

from not at all (10) to very much (100) to judge the items. These item scores were averaged and included in the data analysis. Reliability was also assessed during a pilot study with ten children. These pilot children were comparable to the study participants but did not actually participate in the study. Children completed the experiment twice, two weeks apart. The test-retest coefficients were as follows: task (0.82), ego (0.75), affiliative (0.77), and work avoidant (0.71). Schunk (1996) advised that these coefficients should be reviewed with caution. Some students may have not understood the instrument.

The self efficacy pre-test examined the students' perceived ability to correctly solve 31 pairs of fraction problems. Students answered on a scale of 10 to 100 with 10 being not sure and 100 being really sure. Reliability was also assessed with a pilot test-retest method (r=0.81). students did receive practice time with the scale. They were then shown the fraction problems for two seconds which allowed students time to assess if they could answer the problem without giving adequate time to actually answer the problem.

The skill and persistence pre-test consisted of 31 fraction addition and subtraction problems. These problems consisted of six different categories. These problems were similar to problems students solved during class time (roughly 70% similar). To eliminate any effects of problem familiarity, different forms were used on the pre-test and post-test (pilot study parallel forms r=0.850).

Children were randomly assigned to four conditions: learning goals with self evaluation (LG-SE), learning goal without self evaluation (LG-noSE), performance goal with self evaluation (PG-SE), and performance goal without self evaluation (PG-noSE). Students then attended class sessions for seven days. These class sessions consisted of three phases: a modeled demonstration, guided practice, and independent practice. Teachers received seven instructional packets, one for each session. These packets were standard across the experiment

conditions. Teachers of students in the LG-SE and the LG-noSE conditions stressed the importance of learning to solve problems as opposed to solving the problem. Teachers of student in the PG-SE and the PG-noSE groups stressed the importance of what students were trying to do (performance and skill). These foci were repeated over the seven class sessions.

Product-moment correlations were analyzed with lesson performance and post-test measures to determine relations among relevant variables. According to Schunk (1996), the number of math problems that students completed related positively to self efficacy (r=0.53), skill (r=0.51), and persistence (r=0.42) and negatively to ego orientation (r=-0.50). Self efficacy, skill, and persistence were all positively related with a range of rs=0.63 to 0.89. Task orientation related positively to self efficacy (r=0.48) and skill (r=0.42); ego orientation correlated negatively with self efficacy (r=-0.53) and skill (-0.45). Correlations were also calculated for the self evaluative conditions. The self evaluation score related positively to the number of math problems completed during the lessons (r=0.55). Among the LG-SE students, self evaluation scores correlated positively with post-test self efficacy (r=0.74) and persistence (r=0.77).

The results support the conclusion that self evaluation benefitted student in relation to persistence and motivation students to continue their school work. Just as equally interesting, the number of problems students completed related to a student's persistence, continued interest, and skill level. However, these conclusions should be read with caution due to the low sample number and the low test-retest coefficients of the pilot study for the pre-test and post-test development.

Schunk (1996) conducted a second quantitative study to explore the benefits of giving students a learning goal with or without the chance to self evaluate their capabilities of performance. Study two was a modification of the first study. Participants were once again assigned to a learning goal or performance goal, but all participants completed a self evaluation.

This difference in the study would demonstrate if learning goals were more effective than performance goals or vice versa.

Forty fourth grade students (n=40), 20 girls and 20 boys made up the final participant sample. Once again, these students were from a predominantly middle class background with 21 White students and 19 African American students. These students were seen as average achievers. The same pre-test, class session, post-test, materials, and procedures were used for study two as in study one.

During the first class session, students were given learning or performance goal instructions depending on their condition. The students were then administered a self efficacy for learning assessment by an outside tester, not the teacher. This test consisted of six sample pairs of problems, each representing one of the six class lessons. The students evaluated themselves on their capability to learn how to solve different types of problems as opposed to evaluating how sure they were that they could actually solve the problem. Reliability was assessed in a pilot study with twelve students that did not participate in the final study (r=0.77).

Self evaluation and self satisfaction were assessed at the end of the six class sessions. Self satisfaction assessed the student's satisfaction with their growth in acquiring new skills. For the six sample problems, students judged how happy they were with how much better they were with solving problems compared to how happy they were at the beginning of the class sessions. The students rated their satisfaction on a ten unit scale (10 being not pleased to 100 being really pleased). Students judged their performance with the same ten unit scare (10 being not better and 100 being a whole lot better). Students were asked to think back to the beginning of the class sessions and rate themselves on their progress and growth. Students' goal perceptions were assessed as well. This was done on the seventh day of the class sessions. Students were asked to mark on a ten unit scale (10 = not much to 100 = a whole lot) for four

different items: finish the work, make no errors, learn to solve the problems, and become better in math.

The results of the second study showed that self efficacy for learning was positively correlated to the number of problems solved (r = 0.51) and self evaluation, self satisfaction and learning (range of rs = 0.41 to 0.48). Self evaluation and self satisfaction scores were positively related to post-test self efficacy, skill and task orientation (range of rs = 0.51 to 0.71). Self evaluation was negatively related to finishing the work and ego orientation (range of rs = -0.48 to -0.44). Self satisfaction was positively correlated with learning (r=0.41) and strongly with self evaluation (r = 0.84). Learning correlated with post-test skill, task orientation, and affiliative orientation (range of rs = 0.44 to 0.60).

Based on the results of these two studies, it appeared that giving students a learning goal enhanced their self-efficacy, skill, motivation and task goal orientation. These outcomes are also positively influenced by letting students self assess their performance abilities and growth in gaining new skills. Schunk (1996) provided a theoretical explanation for these results. Schunk (1996) theorized that by emphasizing that the main goal was to learn to solve problems instead of getting the correct answer, students raised their self efficacy for learning and were motivated to notice their task performance and work harder. Self efficacy was substantiated as students observed their progress while acquiring new skills, and a higher self efficacy helped to maintain and increase motivation and performance. The theory is noted but also should be taken with caution as it is not a generalizable conclusion due to the small sample number. Schunk (1996) also noted that self evaluation can also have a negative effect on the student. Sometimes, students with learning problems or students that have not mastered a skill may conclude that they are not capable of learning. Asking such a student to constantly self assess their capabilities may lower, not raise, self efficacy and motivation. This could also lead students to fall into a

cycle of misconceptions in which failure leads to negative self perception, lack of motivation, and more failure. According to Schunk (1996), self assessment should be paired with proper classroom instruction so that students can learn and perceive that they are growing and making progress. Once again, the teacher plays an important roles as a well informed instructor that can fuse instruction, curriculum, and assessment into a meaningful learning experience.

Omelicheva (2005) examined the effects of strong motivation on the reliability of self and peer evaluation as well as the impact of a self and peer evaluation with and without anonymity in a quantitative study. Undergraduate students from freshmen to seniors enrolled in introductory political science courses at Purdue University during the spring and fall semesters of 2003. There were two different experiments. Experiment one had 70 subjects (n1 = 70), experiment two had 40 subjects (n2 = 40).These experiments were embedded in the classroom curriculum. Experiment one used a randomized instrument of evaluation, one with criteria and one without criteria for scoring with anonymous vs. non-anonymous. Students were randomly assigned to one of the experimental groups with no prior training or information. Each student received a folder with an instructions page, a peer evaluation form, one's own essay and an essay of a classmate. The essays were distributed at random. Each student then had to evaluate four peer essays and their own essay. In turn each student was also evaluated by four other students and self-evaluated. The instruction page had the reasoning for participation in self and peer evaluation as well as guidelines on how to do the evaluations. There was not prior training however. Following the evaluations, students filled out an anonymous survey that asked them to rate the contribution of the exercise to their learning. They were also asked to identify ways to improve the exercises and reflect on their concerns.

There were two different peer evaluation forms, one with no criteria and one with guidance on how the assessment should be done. There was also the question of anonymity.

Half of the students received essays marked with classmates' names and the assessors had to sign the evaluation sheet with their name. The other half of the subjects received essays with identification numbers and were asked to sign the evaluation sheet with an assigned ID. Student motivation was also investigated. The instructor promised bonus points to students whose evaluations correlated strongly with the instructors. The other groups were given no such promises for extra points. It was expected that the group with extra points as an extrinsic motivation would have more reliable evaluations.

Omelicheva(2005) focused on the reliability of the peer evaluation process - defined as the extent to which peer assessment contains bias or variable errors. Reliability was measured as the difference between the instructor's evaluation and the peer evaluation. The researcher was also interested in the presence of self bias. This was measured by examining the difference in the self ratings and the instructor's ratings.

Omelicheval composed four different hypotheses: 1) the reliability of peer assessment improves when students are provided with instruments containing definite criteria for evaluation, 2) the reliability of peer assessment improves when students are strongly motivated to apply the criteria to their evaluations, 3) students' self evaluations will be slanted to a higher appraisal of their academic performance, and 4) providing students with criteria for self assessment will increase reliability.

Peer evaluations based on the provided criteria were more reliable than peer evaluations performed with no criteria (M=2.59 with, 2.17 without, p=0.019). Evaluations of anonymous works by anonymous raters turned out to be more reliable (anonymous M=2.5, non anonymous M=2.25) but this was not statistically significant (p=.10). The evaluations for the students from the anonymous evaluation procedure who also used criteria for ranking their peer's essays were not significantly improved relative to evaluations of students from other

conditions. The average self rating was 4.64 and the average peer rating was 4.3 based on a 0 to 5 scale. The instructor's rating was 4.12 and 4.3 which is also significantly different from the 4.64 of the self rating. Ninety-five percent of the students tended to overate themselves by .2 to .5 points. Point five is a whole letter grade. This self bias did not disappear with the different conditions of self and peer evaluations. Students that were strongly motivated to apply criteria when making judgments about academic performance of their peers produced significantly more reliable peer evaluations (motivated M=2.78, not motivated M=2.3, p=0.036).

However, students also voiced their concerns and comments with fellow peers evaluating their work. Students were uneasy about peers being a determining factor in their final grade. They viewed peer assessment as a means to provide each other with positive criticism and advice, and peer assessment could; and as one student claimed, should be used more for helping understand concepts rather than to give out grades. One student even noticed the possibility of evaluation bias due to the controversial nature of the class and the personal positions that some students might hold. Students were very concerned with the evaluator's attitudes and stance toward an opinion task, much like an art assessment task. They claimed that they were not sure if someone could really evaluate an opinion or a feeling. The evaluator may even go so far as to give an extremely low grade if they did not agree with the student that they were grading.

Omelicheva (2005) expressed the concerns that many students have when faced with a peer assessment. Some students may not be at ease with the practice of peer evaluation. There was also the potential negative impact on student learning with peer assessment as used in the formative evaluation of students. Instructors that wish to use peer assessment as a formative assessment should note that some students doubted the constructiveness of peer feedback and warned about student language and the tone of language. These all can lead to negative effects

on the class environment and the confidence of students. There are also numerous biases to address when implementing a peer assessment including gender, racial, and ethnic - just to name a few. Some educators may feel that using peer assessment intrudes on a student's personal information by making it public and readily available to fellow classmates.

Omelicheva's results of her 2005 study appeared valid and very well investigated. However, the students' opinions and attitudes towards peer assessment should not be ignored. There are numerous perils in using peer assessment that the teacher needs to address, prepare for, and ultimately solve to the best of their ability within their classroom.

Hafner and Hafner (2003) conducted a quantitative study to examine the use of rubrics as a valid and reliable tool for peer group assessment of student performance. The top level of rubric communicates what great work should look like and involves the student in constructive learning and self evaluation. According to Brookhart (1999), a rubric begins with a description of the criteria for good work. This checklist should be directly related to the knowledge, critical thinking, or skills that the instructor intended the students to acquire. The criteria should be listed in a descriptive manner instead of listing judgments. The rubric is a simple assessment tool that can and has been used to evaluate and describe levels of performance on a particular task and is used to assess outcomes in a variety of performance based contexts from kindergarten to college. These rubrics should be shared ahead of time so that students better understand what is being asked of them. This involvement can enhance student motivation and learning. In Hafner and Hafner's study (2003), one rubric was developed and used for the same classroom (different students each year) over the course of three years, 1998, 1999, and 2000. Biology students in three different classes ($n1 = 36$, $n2 = 32$, and $n3 = 39$) participated in this study. These students were of a diverse background and were mainly sophomores and juniors majoring in biology. The rubric was used without modification for the entire three year period.

Students worked in pairs of small groups to complete a research project on a topic of interest within the field of human evolutionary biology. The project included library research, critical evaluation on literature, a written interim research report, and an oral and written presentation. The oral presentation was seen as the performance task and was the focus of the rubric. The rubric was created with input from the students (both the original and the modified one that was used). The elements included were organization and research, persuasiveness and logic of argument, collaboration, delivery and grammar, and creativity and originality. These different areas of the rubric seemed to be assessing the product, the process and the students' critical thinking. These were on a point scale with 15, 13, 11, 9, and 7 as the points respectively.

Students were introduced to the rubric at the beginning of the course and encouraged to use the rubric as a guide when thinking about their presentations and projects. The instructor encouraged the students to give honest, confidential feedback and assessments for their peers and that these peer assessments would be considered and evaluated by the instructor as part of the grade. The instructor also asked to see self assessments, but this was not included in their final score. The instructor informed the students of this fact. The instructor also used the rubric to assess student presentations at the same time as the peer assessment, but this was independent of the student rating. Fifty-two of the presentations involving 107 students were evaluated. The data was then entered into a computer with their gender, total points earned in the course, and the year of their class. From this data, the student mean rating was calculated for all students. These values were used to measure the grading rigor of the peer, the mean peer score, and other descriptive states. The student's mean peer score was compared to the instructor's score to determine the validity of this method of assessment.

The total points and the mean peer score in the experiment had no significant difference over the three years (424.722, 434.906, 431.308/ 68.556, 69.22, 69.460). The

instructor also kept consistent scoring in regard to the rubric (69.556, 70.375, 68.556). Absenteeism for the study averaged to approximately 10%. There was almost a 1:1 relationship between the instructor's rating and the student's rating using the rubric. This demonstrated validity in the rubric used for peer assessment. According to student conversations, students found the rubric to be a helpful study tool; and although there was no data on the performance of the class without a rubric, the researchers' opinions were that the overall quality of the presentations was higher and the attendance was better with the rubric than without. This can only be proved by copying this study without a rubric. Results also showed that men and women evaluated each other with the same grading rigor even though they could potentially perform differently when being evaluated. The students recognized consistently higher and lower quality performances as demonstrated by the strong agreement in the ranking order or the presentations for each year. Essentially, the use of a rubric for self evaluation also afforded the students valid assessments of their own performances.

Hafner and Hafner (2003) were very strong in their experimental design, participant sample, and observation period. The only detail that Hafner and Hafner (2003) did not disclose concerned student morale and rigor in assessing their peers' presentations. The researchers failed to explain the total period of time that these presentations were held. Students can lose focus and energy when assessing their peers for long periods of time. There was no information explaining if these participants received breaks or the extent of those breaks during the presentation time.

So what about student perceptions and attitudes toward peer assessment? Keaten and Richardson (1993) asked this same question. In their mixed methods (quantitative and qualitative) study, they focused on four different questions: 1) what are student attitudes towards peer evaluations, 2) do students think that peer assessment is feasible (equitable and

easy to conduct) for project groups, 3) what are student attitudes regarding a specific system of peer assessment, and 4) are student attitudes regarding peer assessment contingent on the dynamics of student work groups?

Keaten and Richardson (1993) collected quantitative data and interviewed 110 undergraduate students from two different speech/communication classes in a mid-western university on their opinions of peer assessment. A peer assessment inventory was developed that consisted of six different areas: attendance at our of class meeting, participation during out of class meetings, attendance at in class meetings, participation during in class meetings, quality of work, and interest in the project. Each of these was assessed on a 10 point scale. The evaluation of peer assessment inventory was developed to examine student attitudes toward peer assessment. This consisted of both a Likert scale and open ended questions such as: allowing students to assess the performance of other students on a group project is fair, the information on the assessment form allowed for an accurate assessment of my group members' contribution, I found rating my peers to be an easy process, I was satisfied with my group's interactions, I was satisfied with the quality of our group's presentation, and I would work with my group again given the chance. Students were also asked to explain their responses to these statements along with rating them on the scale.

Students were assigned to a project group consisting of four to five members at the beginning of the semester. One class was asked to put together a 25 minutes presentation on an aspect of communication (much like the biology class that used rubrics for peer assessment). The other class was asked to do a research project consisting of five steps: making a question/hypothesis, a literature review, collecting data, analyzing results, and conclusions or a reformulation of the question/hypothesis. Both classes were given twelve weeks to complete this assignment. The last three weeks consisted of presentations. At the end of each

presentation, group members were given a copy of the peer assessment inventory. They were asked to rate all members. Responses were kept confidential and students were to not discuss with the other group members. In the final week, students were given the evaluation of the peer assessment inventory. The results of the evaluations were as follows: 88% found peer assessment fair, 79% thought the form was accurate, 67% thought rating peers was easy, 67% were satisfied with their group interaction, 65% were satisfied with the group presentation, and 54% said they would work with their group again.

A correlation matrix was set up to determine relationships. Attitudes toward group interaction were positively correlated with attitudes toward working with the group again ($r = 0.76$, $p<0.01$). Attitudes toward group interaction were positively correlated with attitudes toward the quality of the group's presentation ($r = 0.47$, $p<0.01$). Attitudes toward the quality of the group's presentation were positively correlated with attitudes regarding the willingness to work with the group again ($r = 0.46$, $p<0.01$). Attitudes toward the fairness of peer assessment were positively correlated with attitudes regarding the accuracy of the peer assessment inventory ($r = 0.43$, $p<0.01$).

The researchers created four categories for participants to respond to the accuracy and fairness of peer assessment: individual strengths, contributions, accountability, and responsibility. Of the 93 qualitative, narrative responses, only four were negative in regard to performance in group dynamics and personality. The main frustrations focused on noncontributing group members and no clear reference or criteria to assess peers. There were categories that the participants wished to add as well: student commitment to projects outside of group meetings (both in and out of class) and sensitivity to subject bias. They also encouraged an area for narrative responses and a means of formative assessment through the process. Students also viewed per assessment as a necessary part of the evaluation process and it

allowed students a chance to develop evaluation skills.

Keaten and Richardson (1993) have presented a clear case for peer assessment as an effective method of assessing peer work. However, the sample size is quite small and the steps of the experiment design are not clearly stated. The participants were part of a very specific content area that could possibly be more interested in human interaction as it was a communication course. The visual arts are seen as a means of communication as well; however, speech communication and visual art communication differ in many ways. This could influence the group dynamics and participant views on peer assessment. The researcher also did not test the validity or reliability of the peer assessment inventory. This should be done in future studies of this fashion. The method used for peer assessment only consisted of a peer review. There are other means of assessing peers as well, like performance assessment for example. The researcher lists peer review, peer nomination, and peer ratings as three different methods that could be studied in the future.

The study's results also could have been affected by the teacher and student familiarity with this assessment method. If the teachers and students had received more training and guidance, the results could have been substantially different. Other studies of this nature have included assessment method training as part of the experiment design to ensure that all participants and test or assessment administrators are familiar with the design and purpose of the assessment method.

Tanner and Jones (1994) conducted a qualitative study to investigate the importance of peer and self assessment in the development of a student's modeling skills in a Welsh mathematics classroom over the course of a single school year. In 1991-1992, the Welsh Office funded the Use and Practical Applications of Mathematics Project to develop approaches and materials to teach and assess the cognitive skills used and applied in mathematics in modeled

situations. Along with this variety of teaching methods, a variety of assessment methods were developed such as peer and self assessment that was facilitated by the teacher. The students rehearsed negotiating classroom values and the metacognitive skills of planning, monitoring, and evaluating. Mathematical modeling employs a real world problem in mathematics with a basic framework for solving that problem. According to Tanner and Jones (1994), knowledge alone is not sufficient for modeling to be successful. Students must also choose to use that knowledge and understand the progress that is made. There are three strands of metacognition that can support the modeling process. Those strands are planning, monitoring, and evaluating. During the modeling process, mathematicians cycle and alternate between those strands, suggesting new ideas, evaluating those ideas, and monitoring the changes to come to more new ideas.

Eight different secondary schools participated in this study. Two teachers from each school attended a course on practical mathematics. This course was to introduce the teachers to mathematical modeling and allowed a space for teachers to develop the activities they would be doing in the classroom. These two teachers then returned to their respective schools to inform their colleagues of this process. The research did not specify exactly how many teachers in total participated in this study.

The school lessons were monitored with self evaluation forms and by university researchers. These university researchers acted as participant observers of 100 lessons. They also made their purpose known to the study participants and students involved as they stopped students to ask questions and discuss the modeling process students were doing. Researchers collected data by taking written notes, video, and audio recording. Students were questioned and recorded during the activities to determine their perceptions and strategies. Informal interviews with the teachers and selected groups of students were also recorded after the

lessons. The teacher researchers also participated in regular meetings to discuss their experiences with the activities and to develop better teaching strategies, assessment methods, and tasks. These meetings were also tape recorded. These tasks were created to be practical tasks for real world situations and problems. Students formulated their own mathematical models. These tasks were completed by the whole class, spanning all ability ranges. The tasks also extended over the course of several class periods and a variety of equipment was available to the students.

Tanner and Jones (1994) used the student observations and interviews to find the students' perceptions and attitudes of their situation and the strategies they were using to find a solution. Instead of observing whether students failed or succeeded, the researchers recorded how much help students needed to make progress. They assessed the development of the students and their modeling skills. This assessed the whole unit of the social system as opposed to the individual student. Task specific criteria were developed from the observed behaviors. These were documented in three strands with a ten point scale: application/strategy, logic, and communication. Behaviors that had been consistently observed in average ability in 14 year old students equaled a six on the scale. For the assessment, the criteria were written in student friendly language. These criteria were discussed and modified during the teacher meetings. These criteria were not a perfect match with the national curriculum, but they did fit very closely. These assessment frameworks proved to be quite easy for teachers to use when evaluating student work. They were also reliable in that the grades between teachers were generally consistent. There is no quantitative data to support this however. The assessment of work did lack validity due to the different class presentation methods between teachers. For example, for some teachers, there was so much structure in the modeling process that the students were not able to fully demonstrate their performance. On the other end, some

teachers were unwilling to intervene with the modeling process which left students unsure of expectations resulting in student work with no apparent structure. In future studies, a standard structure should be employed or contextual evidence could be provided to ensure validity.

Early in the experiment, students failed to report to their achievements, possibly because students did not completely understand what was considered important for their work. They had not been accultured to the modeling process. This resulted in low grades. This also proved that the process of acculturating students into the modeling process and developing peer assessment skills was crucial. Tanner and Jones (1994) stated that assessment is integral to acculturating students. In order for students to gain access to a classroom, assessment is necessary for student understanding. Tanner and Jones (1994) suggested informal formative assessments to give the student a chance to negotiate ways of thinking and behavior that are acceptable for the teacher. Based on their study, Tanner and Jones (1994) suggested several teaching methods that utilize peer assessment skills for the classroom. These methods included using self monitoring and reflection to evaluate progress while working on each state of the problem, presenting scientific findings and approaches to the class for discussion and questioning, and encouraging students to look back at their work as a group for better strategies to problem solving.

Students were also encouraged to self assess their work based on the frameworks. This self assessment can create a dialogue between student and teacher. Student self assessment varied in accuracy from class to class. According to the researchers, self assessment involves both an understanding of problem solving and reflection about the process. This was absent from some classrooms. The students that were not familiar or experienced with self assessment rated themselves higher than their teachers; a theme in several studies that examined the validity of self assessment. However, Tanner and Jones (1994) saw this as an opportunity for

more student/teacher discussion and socialization as the student and the teacher negotiated their ratings.

The researcher observations also noted that students were more successful in self assessment if they were first involved in peer assessment by experiencing the peer assessment. Students were also to reflect on what was important in their own work.

Tanner and Jones's (1994) conclusions appear to be valid and worthwhile based on their experiment methods. Their use of triangulation in collecting data as well as their extended observation time allowed the researchers to collect very rich and important information concerning student assessment and peer assessment and their relationship to teaching instruction and student assessment training. Tanner and Jones (1994) believed that assessment should aid instruction by giving the teacher a look at the development and progress of a student. It should also act as a means of open communication between the student and the teacher. Peer and self assessment provide a means for students to negotiate a good process and solution to a real work problem. Peer assessment requires students to discuss, challenge and prove their processes and products. Through self assessment, the student develops the awareness for the need for planning, monitoring, and evaluation. By participating in peer and self assessment, the student is involved in a cyclical learning process that supports the development of metacognitive skills (Tanner and Jones, 1994).

Summary

Self and Peer assessment have been found to be valid forms of alternative assessment. However, these assessment methods must be practiced to ensure that students are rating one another and themselves reliably and fairly. Students voiced concerns of student bias, misunderstandings, or grievances that may have an effect on how a peer grades another peer. Once again, these forms of alternative assessment need to be set in a safe environment, in a correct curriculum with proper instruction.

The following section reviews and analyzes the research on performance assessment as an alternative means of evaluating students.

Performance Assessment

Bergee (1997) conducted a quantitative study to examine relationships among faculty, peer, and self evaluations of applied music performances. Bergee (1997) asked what is the inter-judge reliability of faculty and peer evaluations of performances and to what extent do faculty, peer, and self evaluations of performances correlate.

Bergee (1997) asked music graduate students and faculty members of a large Midwestern university to participate in evaluating undergraduate voice, percussion, brass, woodwind, and stringed instrument performances. The final number of faculty that did participate totaled 19 assessors (n = 19). The undergraduate participants were chosen at random. These included seven percussionists, seven vocalists, eight brass instrumentalists, nine woodwind instrumentalists, and six stringed instrumentalists. These participants were of varying background, academic level, and performance achievement level. Additional faculty members and undergraduate students participated in the study from a small private college (three faculty members and five students) and from a regional public university (one faculty member and eight

students).

The performances evaluated were nine minutes in length and videotaped so that multiple copies could be made and sent to different assessors. These tapes were erased at the end of the evaluation. Performance outcomes were quantified through a standard criteria found in the Music Educators National Conference solo adjudication forms. Students received a videotape of the performances at their universities with copies of the scoring forms ten days after the live performances. They were to rate the performances as accurately and objectively as possible. These performances were randomly given to the assessors by randomly ordering the videotapes three times, re-recording the performances, and then randomly pairing student assessors with one of the three performance orders. The faculty from the universities evaluated two randomly chosen performers that were not from their own universities to avoid the effects of familiarity.

Faculty total score inter-judge reliabilities were acceptable, ranging from 0.74 to 0.93. Category inter-judge reliability was consistent as well, ranging from 0.70 to 0.90. Peer total score reliability ranged from 0.83 to 0.89. The range of inter-reliability between peers and faculty was very close, falling between 0.70 and 0.90.

Self evaluations, however, were another story. Self evaluations correlated poorly with faculty and peer evaluations for both total scores and category scores. Only eight out of 92 correlations were higher than 0.50 with 46 of the correlations being negative. No pattern of self evaluation was higher or lower than faculty evaluations or peer evaluations. The total and category scores of the faculty scores were generally uneven in inter-judge reliability compared to the student peer inter-judge reliability which was more uniform than the faculty's.

Bergee (1997) was concerned with the wide range of faculty inter-judge reliabilities. In two of the three participating universities, Bergee (1997) noted faculty unreliability. There was a

theoretical explanation for this. Regular assessment of performance skills is crucial to the development of music students and teacher. Therefore, music students and teachers should have regular access to regular, reliable, and valid performance assessment. Creating performance assessment standards could combat the unreliability found in faculty evaluations, especially for faculty member with limited experience assessing student performances.

Nadeau, Richard, and Godbout (2008) conducted a quantitative study to determine the effects of performance assessment from a physical perspective. Nadeau, Richard, and Godbout (2008) examined the effects of peer assessment as an assessment of a performance task by investigating if a team sport assessment procedure could be a reliable and valid mean of assessing hockey players. Nineteen bantam and midget players between the ages of 14 and 17 from different elite teams in the Quebec City area of Canada participated in this study. The assessment used in this study was to determine four different aspects of performance, technique, tactics or decision making, product (the end result) and process (how did you do that). Hockey was typically assessed in a standardized form, only focusing on one of these aspects, the technical product.

Two coaches from the sports studies program in Quebec area acted as experts to verify the assessment strategies and its components and variables. These experts had 5 to 15 years experience as physical educators. These two coaches categorized the player into different skill levels prior to the data collection by observing a scrimmage hockey game. These categories were used against the student's observations for concurrent validity.

Players were divided up into teams of three each player had to observe two other players and all players were represented by numbers. They used an observation rubric which they received training for prior to their games. This training lasted for 35 minutes and 20 minutes in two different sessions. All observation times were two minutes total to reflect the

intensity and “normal” time frame for a hockey player to be in the game. The researchers tried to make the observation time as close to game time as possible. Goal tenders were not assessed during these observation times however. The two different peer observations for one player were compared for inter-observer reliability. Levels of agreement ranged from 59% to 95% with only two being less than 78%. Well trained adult observers were also used and their level of agreement ranged from 80% to 82%.

In essence, the players were participating in a means of self and peer assessment. Their peers were assessing their skills, both technical and decision making. The players then used this assessment and evaluated their performance as well to see where they needed further practice or development. This was more of a formative assessment as opposed to a summative assessment.

In the concurrent validity test, there was a correlation of -0.77, the negative is due to a player’s high performance score in relation to a low rank score. A correlation of 0.70 or more indicates validity. It is safe to say that this means of assessment is reliable in determining player performance.

Nadeau, Richard, and Godbout (2008) presented a very compelling case for the use of peer assessment as a means to assess a performance task. However, this conclusion is difficult to generalize or transfer to visual art students due to its very small sample size and the short research observation time. It was also noted that observing hockey players is a very difficult task for those that are new to the assessment process. The participants may have performed and assessed their peers differently with more practice observing hockey players in a read game situation. The rules of the game of hockey were also changed for the purposes of the study. This could have affected the authenticity of the assessment and the authenticity of a player’s performance. For example, in the game observed, a main rule had changed. Instead of a mid-ice

face-off after a goal, as is customary in a real hockey game, the defensive team had to start behind their goal. This could have affected the players' performances and observations.

Richard, Godbout, Tousignant, and Grehaigne (1999) conducted a qualitative study to determine a means of observing and rating performance behaviors that showed a student's ability to problem solve by making decisions, moving in the correct manner, and executing the proper skills needed in a basketball game. The assessment process was based on two notions: how a player gets the ball and how a player disposes of the ball. This style of assessment quantified a player's overall offensive performance. This reflected technical and tactical skills of players. Another characteristic of this assessment process involved the collection of data. In order to make this assessment authentic, students as active assessment participants were included in assessment for the teaching/learning process.

Six elementary school physical educators (n=6) from southeast New Brunswick participated in the study. The researchers did not specify how may classes these educators had. Teachers used the assessment process for two five minute periods for grade 5-6 and two seven minute periods for grades 7-8. The teachers were trained for the assessment prior to this by attending two 90 minutes workshops. During the assessment periods, students collected data and informed players of their results.

Each teacher was interviewed for one hour. The interview was a standardized, open ended interview that was conducted by one researcher and audio recorded. Two themes and eleven questions came up from these interviews. The two themes were pedagogical uses and implications and different practical issues related to preparing students for observing and collecting data. Teachers received an interview guide a week before their individual interview to reflect and prepare their answers and questions.

The transcribed data from the teacher interviews was reduced down to four main

categories and twelve subcategories based on the two themes and questions. The four categories were development of students' tactical awareness, integration of the procedure in regular assessment data practices, instruction of students concerning the use of the procedure, and the collection of assessment data. All of the teachers stated that this assessment procedure stimulated the students' interest in learning more about their performance and how to improve their performance. Older students were observed to benefit more from this assessment process than the younger students, linking their results to their performance improvement. This assessment model also provided opportunities for students to identify their strengths and weaknesses in their performances and make comparisons in relation to their peers and the performance criteria. Students reflected on their performance and this reflection process made them more aware on the court.

One researcher noted that the assessment gave instant, objective feedback to the students and that the assessment was very adaptable to a physical education curriculum. Even though some of the students had difficulty observing their peers, all teachers stated that they were confident in the accuracy of the data collected by their students. The average precision correlation for the volume of play performance varied between 0.95 and 0.97 ($r = 0.95$ to 0.97) for all the grade levels involved. The average precision correlation for the efficiency ranged from 0.88 to 0.90 ($r = 0.88$ to 0.90). Inter-observer reliability was also established for both the volume of play and efficiency index. Volume of play coefficients exceeded 0.80 for all classes. Efficiency correlations were at 0.75 for all grade levels except for grade five which had an efficiency correlation of 0.65. This number went up to 0.75 when three students were not included in the data. The teachers had good reason to be confident. Based on this data, one could say that in an average class size of thirty students, three students will have difficulties completing a peer assessment.

Some of the problems that teachers came across while doing this performance assessment model involved student comprehension of variables and time. Students had a difficult time defining certain observations in the assessment. Students also lost motivation while the teachers instructed them on how to properly use the observation worksheets. Students were used to being active in physical education classes, much like students are used to making art in an art classroom.

It may be premature to draw conclusions from this example of physical education performance assessment as well. The sample number was very small as well as the time the researchers spent in training and preparing the teachers and the students to participate in this study. However, by combining these findings with those of Nadeau, Richards, and Godbout (2008), a trend emerges regarding the proper ways to implement peer assessment for a performance task. Students are showing positive correlations when tested for inter-reliability which would lead one to conclude that peer assessment is a reliable means to assess performance. However, only peer assessment, with no other assessment methods, does not seem like a reliable or motivating means of evaluating student growth. Instead, peer assessment seems to be more valid for peer support and for providing objective, immediate feedback that can help a student self assess and reflect on improvements in their performances.

Fuchs, Fuchs, Karns, Hamlett, and Katzaforr (1999) conducted a quantitative study to investigate the effects of classroom based performance assessment and its corresponding instruction practice on student problem solving. The alternative form of assessment examined in this study was a response to problems associated with traditional standardized tests. Sixteen general educators (n=16) from four different schools in a southeastern urban school district were all randomly assigned to an instruction condition, half to a classroom based performance assessment driven instruction. Student participants that were included in the final data analysis

consisted of the students that were present (no school absences) for the entire study (n=272). Teachers were asked to review their student's current academic achievement status and designate their students as above, at, or below average.

A standardized pre-test on computation and application mathematics was given to the student participants. Six different parallel forms of the performance assessment (PA) were developed for each grade level. The researcher held teacher meetings in which the teachers had to individually complete the performance assessment, learn to use the PA rubric, and divide into grade levels teams to modify the PA rubric. This PA and rubric was piloted with three students who were entering and three students who were exiting the targeted grades. This revision process recurred three times. In regard to reliability and validity, the PA was investigated using three criteria: inter-task consistency, error due to raters, and relation with other measures. PAs proved to be an adequate assessment although no numbers were given.

During the first week, teachers participated in a full day workshop to learn about the reform emphasis on excellence and equity and about the purpose of the PA, complete one PA, review scoring criteria, achieve scoring reliability, and discuss possible instructional methods. Teachers administered the PAs on the Monday of school weeks 13, 13, and 23. The Tuesday after each administration, the teachers came back together to review scoring procedures, achieve scoring reliability, and discuss possible instructional methods. The following Wednesday or Thursday, teachers distributed the scored performance assessments and their feedback to students.

Teachers completed a two part questionnaire. For one part, teachers responded to open ended questions to assess their knowledge of what a PA was and how a PA might improve their instructional design. For the second part of the questionnaire, teachers distributed 100 points to indicate how much instructional time they devoted to basic math facts, computation, work

problems, problem solving activities, and other activities. On release days, teachers also completed an instructional plan sheet that described the classroom activities that she planned to use for the next performance assessment to enhance student performance on the assessment. Student problem solving was also measured using three types of measures: analogous, related, and novel with respect to the performance assessments. The analogous measure involved three unused and unseen parallel forms of the performance assessment that had been developed for the student's grade. The related measure involved one unseen performance assessment from a grade level below the student's grade. This incorporated a different problem structure and required a different set of applications from the analogous measure. The novel measure was a basic skills test.

PA teachers' problems with performance assessment represented an average of 5.13 with the no-PA teachers; problems reflecting an average of 2.25 dimensions. These results indicated that classroom based performance assessment driven instruction did increase the teacher's knowledge of what a performance assessment was. Teachers were also asked to explain the ways that a performance assessment could be helpful in making mathematical instruction decisions. PA teachers reported a total of 3.25 while no-PA teachers reported a total of 1.75. On the questionnaire that asked teachers to reflect the amount of instructional time they spent on a mathematical skill, there was a significant interaction between treatment and year. PA teachers decreased their instructional time on math facts (last M=15.71, this M=13.75) and computation (last M=25.00, this M=18.75) while increasing time and emphasis on problem solving activities (last M=6.43, this M=15.00). The no-PA teachers did not show much sign of an change on emphasis (math facts: last M=17.50, this M=20.00, computation: last M=20.63, this M=21.88, problem solving: last M=8.13, this M=6.88). This comparison between PA teachers and no-PA teachers shows a shift in curriculum focus away from basic math facts and computation

to problem solving for those teachers that used performance assessments.

In relation to the student achievement level, across PA conditions, scores of the above grade level students were higher than those at grade level, whose scores were higher than students below grade level. Growth for above grade level students was greater in the PA condition (comp M=4.00, ps M=3.83) than in the no-PA condition (comp M=1.83, ps M=1.78). Growth for at level students was also greater in the PA condition (comp M=2.66, ps M=2.26) than in the no-PA condition (comp M=1.24, ps M=0.98). Growth for below grade level students was comparable in the PA condition (comp M= 1.35, ps M=1.05) and the no-PA condition (comp M=1.19, ps M=0.71). With the teachers use of the PA driven instruction, it appears that the above and at level student increased their performance. Results indicated that the classroom based performance assessment instruction increased the teachers' knowledge of what a performance assessment is and its purpose. It can also move student toward more problem solving thinking skills as opposed to recall and basic computation.

These teachers used the performance assessment to coach students. This could lead to increased scores in the performance assessments without representing any really learning however. This study further demonstrated the point that teacher instruction and comprehension of a performance assessment can influence a student's learning. The teachers were very well organized and diligent in creating, questioning, revising, and modifying the forms of the performance assessment. This helped the teachers to better understand the purpose of the assessment and better help instruct their students on the assessment procedure. The ideas these teachers brainstormed contained three important instructional strategies: incorporating activities that are designed to expand the problem solving ability of the students, giving students a chance to discover relationships in their knowledge by extending the problem solving activities, and helping students demonstrate and communicate the competence they already

possessed (Fuchs, Fuchs, Karns, Hamlett, and Katzaroff, 1999).

Baxter, Shavelson, Goldman, and Pine (1992) assessed the scoring system for a performance assessment for a hands-on science experiment with fifth grade students in their quantitative study. Two groups of fifth grade students from two different school districts participated in this study. The two schools had different levels of hands on experience in the science classroom. There were 41 students (ES, n=41) from the school district with experience and 55 (IS, n=55) students from the school district with very little hands on experience. The inexperienced students were from a district that adhered to a strict textbook policy. These subjects completed laboratory experiments to determine which of three different paper towels was the most absorbent.

Prior to this hands on experiment, students took a multiple choice test to assess their basic science skills. This test served as the traditional means to assess student performance. For the hands on experiment, student had to choose a method to get the paper towels wet, decide whether or not to completely saturate the towels. And choose a method to determine their results. A rubric in the form of a flow chart was created to judge the students' decisions in the experiment. If students did not follow a logical science path, or if a step of the scientific process was left out, the letter grade was changed accordingly. There was more than one possible solution to this scientific problem however and a different path with the right number of procedures could result in the same letter grade.

There were eight research observers that were trained for the experiment. The observers had to complete the paper towel experiment as well. The score system was then explained with reference to the observer's methods in solving the paper towel problem. They then viewed, scored, and discussed three videotapes of students doing the paper towel experiment. The observers evaluated a range of student performances and asked questions

regarding interpretation of the scoring system. After the discussion, the observers then independently viewed and scored six videotapes. Observers' scores were compared and a person-by-observer generalizability coefficient of 0.94 suggested that observers rated student performance similarly. The researchers examined validity in three ways: ES-IS student comparisons of scores for both the multiple choice and performance assessments, the relationship between multiple choice scores and performance assessment scores, and the relationship among ability, multiple choice scores and performance assessment scores.

Students also kept a notebook to record their observations of the performance assessment. This was used to determine the correlation between the scores based on the hands on experiment and the procedures noted by the trained observers. The notebooks did a fairly good job showing the performance of students regardless of experience (ES r=0.91, IS r=0.82).

These results suggest that a hands on experiment can be scored reliably. This study also noted that the experiment can be carried out without specific scientific knowledge. The performance correlated less with ability (ES r=0.21, IS r=0.51) than did the multiple choice test (ES r=0.61, $p<0.05$, IS r=0.70). This study did not mention different teaching strategies in implementing this type of performance assessment. There was a difference between the performance scores of the experienced and the inexperienced students which may suggest that prolonged training with a performance assessment helps students become familiar and comfortable with performance assessment. In regard to performance assessment as a reliable means of assessment, the low sample number makes generalizability difficult. This study in conjunction with another study of the same nature could produce more worthwhile conclusions.

Lane, Lui, Ankenmann, and Stone (1996) conducted a quantitative experiment to determine the validity and generalizability of cognitive assessment instrument designed to measure the outcomes and growth of middle school students. Inter-task and inter-rater

consistency were examined to investigate the generalizability of the test scores as well.

The QUASAR Cognitive Assessment Instrument (QCAI) measures the outcomes and growth in math and helps determine the success of middle school programs that promote mathematical thinking and reasoning skills. The sixth grade test consisted of 36 open ended tasks that involve problem solving, reasoning and communication. A general holistic rubric was used for scoring student responses. The rubric incorporated three inter-related components: mathematical conceptual and procedural knowledge, strategic knowledge, and mathematical communication. Student responses were rated by middle school teachers.

The QCAI was administered for one class period. Four different forms were randomly given out within each sixth and seventh grade class in six participating schools. The sample size consisted of 1822 sixth graders (n6=1822) and 1782 seventh graders (n7=1782). These students varied in ethnic, racial, and socioeconomic background. Before the test, the teachers were asked to explain the degree to which the math content and processes assessed by the tasks were consistent with the goals of their instruction.

Three types of generalizability designs were used to examine inter-task and inter-rate consistency: person by task (p x t), person by task by rater (p x t x r), and person nested within a school by task ((p:s) x t). The p x t design was used to examine different student performances across different tasks. The p x t x r was used to examine the errors in measurement dues to the unreliability of raters. The (p:s x t) design was used to examine if the generalizations were valid.

The generalizability coefficients ranged from 0.67 to 0.83 when n=8 and from 0.71 to 0.94 when n=9. The dependability coefficients ranged from 0.67 to 0.80 with n=8 and from 0.69 to 0.82 when n=9. These coefficients were consistent with the generalizability reports from administration the year before. The generalizability coefficients ranged from 0.74 to 0.84 and the dependability coefficients ranged from 0.71 to 0.80. Based on these results, it can be

determined that this method of performance assessment testing in mathematics is both valid, reliable, and generalizable across different student backgrounds.

Greer (2001) conducted a mixed methods study to determine the impact of changing instruction and assessment methods of a performance task on the overall performance of undergraduate accounting students. One hundred and seven students completed the end of semester exam in 1999 and 110 students in the in course assessment. Different students from different classes, such as Accounting and Finance, Accounting with Law, Business Studies, and Business Management studied this module.

Triangulation was used to collect data. Quantitative difference in learning outcomes were based on the students' assessment marks on the module. Qualitative information of staff perceptions of the students' performance came from feedback from tutors. Students evaluated their learning environment by responding to a teacher evaluation questionnaire and through informal discussions with students in a focus group. This focus group was started seven weeks after the end of the semester. This was done to determine the perceptions of a sample of students on the module. Six students volunteered for this group and attended the meeting. At this time, students had been given their results and had time to reflect on their first semester experience. They gave honest answers and researchers assured the students that their answers would be used only for research purposes.

Students were given an assessment criteria when the in course assessment was given out. Students were required to prepare financial statements from the provided data and also to write a professional letter to a mock client to answer client questions. A critical evaluation of the information required students to really think about their responses and to analyze the data. When the assessment was returned, students were given a complete feedback form from the assessor. This identified how students performed against the criteria and gave written

comments on how it could be improved. This feedback form allowed for the maximum amount of feedback in the minimum amount of time. The focus group felt that the feedback form helped to indicate areas of weakness and areas for improvement.

Compared to the prior year's assessment model, a majority of students stated that the in course assessment helped them understand the content of the module better. This assessment was seen as a means to facilitate learning, not just as a letter grade. The focus group stated that the assessment module forced students to read and work hard – both things needed in a good examination. This assessment model helped students to develop a better understanding of the content of the module according to the focus group (Greer, 2001). In comparing the prior years to the present year of the study, it was difficult to determine if there had been a difference in understanding as the assessment and student history of the previous year was not known. Although no real conclusions can be pulled from this study due to the lack in comparison between the previous years and the study year, the qualitative information is helpful. The statements of the student focus group agree with many of the statements of other qualitative performance assessment studies. There are several areas that the educator needs to provide in the environment, the curriculum, and the assessment that allow for deep learning to take place.

Summary

Performance assessments were also seen as a valid assessment that students found engaging and helpful in their growth and development. Performance assessments can also be reliably judged by outside judges, teachers, and students. This assessment method should also be properly practiced and introduced. Performance assessment can provide students with immediate feedback and can help students to see areas for improvement in their performance

and their thinking or skills.

The next section aims to analyze and examine the affects of alternative assessment on both teacher and student. This section with also examine the effect teacher attitudes can have on student performance and completion of an alternative assessment method.

Effects of Assessment on Students and Teachers

In the previous sections, different methods of assessment have been examined including portfolios, self assessment, peer assessment and performance assessment. This section attempts to examine the effects that these types of assessment have on both students and teachers.

A quantitative study by Dochy, Gielen, Janssens, Schelfhout, and Struyven (2006) investigated the effects of end of course assessment on student performance by comparing multiple choice testing, peer assessment, case based assessment and portfolio assessment. They hypothesized that students who learn for understanding may actually perform better than students that learn for knowledge acquisition or memorization. They went on to form three more hypotheses: 1) if standardized tests measure mainly knowledge construction, the overall scores of the student engaged in peer, portfolio and case based assessments will be higher than the scores of the multiple choice students, 2) if traditional assessment serves knowledge acquisition purposes and alternative assessments aim at knowledge construction, the multiple choice students should do as well as students using peer, portfolio, or case based assessments and 3) if the characteristics of the learning environment are taken into account, the active knowledge construction students will do better before preparing for the exam than their peers because the first group interacts more with the learning materials.

Eight hundred and sixteen undergraduate students (n=816) in eight different teacher

education programs participated in this study. Multiple choice questions were designed to encourage student learning through providing annotations. Case based assessment is seen to be the most effective method for measuring student learning according to different business programs. These are authentic cases and problem solving assignments that give evidence of student learning and understanding. Peer assessment commented on others to help students improve their own presentation. Portfolios were an alternative method of assessment that students tended to like. Students thought of portfolios as helpful in learning outcomes but little evidence was available of the effect of portfolio assessment on student performance.

Five different institutions were given five different research conditions. One school instructed their students through a lecture based learning environment. They were given formal lectures and assessed by multiple choice exams. The other four groups learned in a student activating learning environment. Each was assessed though different means - case based, peer, portfolio, and multiple choice. Multiple choice participants were given 20 questions with four answer options with only once correct answer. To avoid guessing, wrong answers resulted in a deduction of 1/3 point.

Case based students were allowed to use all resources (school reports of the class, the week planning, the thematic planning of the whole school year, the floor plan, a medical report, and a letter from a mother. All the students received the same documents and used these for their exam.

Peer assessment subjects were given assignments on child development which required students to work in groups of six to eight. Each student in the group scored their peers and themselves on the processes within the group and the contribution of each member. Essentially, did students contribute better, equal or less than the average? The teacher evaluated the end product and the peer assessment was applied to the final score.

Portfolio assessment consisted of a selection of the assignments with student reflections on the experiences. A pre-test and post-test were used to find out whether each of the four different assessment methods had differentiating effects on student learning test one (n=534) measured what a student learned and remembered about the course after experiencing the learning environment. Test 2 (n=552) assessed student performance after preparing for the assessment method. This was administered right after their exam or assessment.

The data did not support the first hypothesis. The lecture multiple choice condition scored very well on both test 1 (M=6.32) and 2 (M=5.93). There are two theories for these conclusions: the tests in this study did not mainly measure student understanding and therefore the scores among the assessment methods did not differ or multiple choice students also study for understanding. Perhaps the characteristics of the tests in this study advantage students from the multiple choice conditions during the post-test. In order to know for sure, one would have to examine the study habits of students and their investment.

Multiple choice conditions outperformed the other conditions for almost all the four categories of knowledge processing: knowledge acquisition (0.61), insight to knowledge (0.84), skills to apply (0.51) and problem solving skills (0.67) except for problem solving skills. The other conditions varied in their scores as follows: active multiple choice - knowledge acquisition (0.54), insight to knowledge (0.80), skills to apply (0.45) and problem solving skills (0.59); case based - knowledge acquisition (0.42), insight to knowledge (0.72), skills to apply (0.32) and problem solving skills (0.57); peer - knowledge acquisition (0.49), insight to knowledge (0.68), skills to apply (0.42) and problem solving skills (0.66); and portfolio - knowledge acquisition (0.45), insight to knowledge (0.69), skills to apply (0.43) and problem solving skills (0.71).

End of course evaluation methods and more specifically the requirements take a difference. There were slight differences in student scores in both tests. The portfolio

assessment required a considerable amount of last dig effort when test 1 came up, which resulted in high performance at the end of the lesson. The multiple choice exam required that students master a huge amount of content that was not present taking the exam since it was closed book as compared to the other assessment methods that allowed resources. This was an important limitation to note. The researchers recommended that a triangulation of assessment formats in necessary to compare student performances in future studies.

Assessments may not produce overall effects on student learning because not all assessments in this study obtained comparable results. There were more findings on the inconclusiveness of the study than actual results.

The quantitative study by Gerbert (1986) asked four questions in regard to the effect evaluation has on student motivation and learning: 1) what effect do evaluative statements have on artistic quality as revealed in art products, 2) what effect do evaluative statements have on amount of time spent on tasks, 3) what effect do evaluative statements have on student desire to do more work (continuing motivation), and 4) what differences in art performance and continuing motivation may be attributed to such things as gender, attitude and creative ability.

Eighty fourth grade students (n=80) from an Indiana county school group were chosen for this study. This particular school system had eight elementary schools that range from rural to urban and serve varying socioeconomic background. Twenty subjects (10 male and 10 female) were randomly selected from each of four different schools. These groups were assigned to random evaluation conditions – teacher, peer comparison, self evaluation, and a control. Fourth grade students were chosen due to the research that stated that students at this age have a decrease in motivation toward school subjects including art. It is also thought that at this age the student also thinks that art ability is unchangeable.

Three different pre-assessments were used to test for group homogeneity in the groups.

Art attitude was measured by the Art and Me scale, a 12 item semantic differential scale (r=0.896). Creative potential was measured by the group inventory for finding creative talent which consists of 34 yes/no statements (r=0.86). Figural drawing ability was measured by the Torrance Test of Creative Thinking, Figural Form A, which included three drawing tasks: completing objects or pictures when given a shape, adding lines to 10 incomplete figures, and adding 30 sets of lines to make objects of picture drawings, titling each. The reliability of this figural drawing test varies from 0.35 to 0.75, but it is one of the few available standardized tests on figure drawing that is normed.

Students were asked to do two different tasks. First, they were asked to draw an animal of their choice trying to show where the animal was and what the animal was doing using eight crayons and white 8.5x11 inch paper. They were also asked to make their images as big as possible to take up as much space on the paper. The second task was a tangram task. Students were asked to use seven geometric cut out shapes to create an object, person, or interesting arrangement. These were then glued onto 8.5x11 inch paper.

The different evaluation conditions were teacher evaluation (subjects were told that their results were important and would be graded), self evaluation (subjects were told that they alone would evaluate their art performance), and peer comparison (subjects were told that their art performance would be rated by an investigator on how their work compared to others). Two trained judges used a five point scale with five criteria (spatial, expressive content, detail, color, and skill) to rate the design quality of the drawing and tangrams. The judges were blind to subject gender and nature of evaluation condition of the environment.

After the tasks, the subjects were asked to respond to three sequential questions. The first involved subject indicating their eagerness to continue. Next, the subjects had to state if they do such tasks on their own time. They then, if answering yes had to indicate a specific time

and sign the sheet. Their responses were scored using a six point scale.

The mean scores for the Art and Me Scale were 43.95 for the teacher group, 46.15 for the peer group, 44.25 for the self group and 43.95 for the control group. This means there were no significant differences in attitude toward art among the groups. The mean scores for the Group Inventory for Finding Creative Talent were 23.50 for the teacher group, 21.35 for peer, 24.40 for self and 20.35 for the control. This showed that there was a slight significant difference between boys in the control and self groups. The other groups and females were essentially the same. An analysis of the mean scores for the Torrance Test of Creative Thinking, Figural Form A revealed no significant difference between the groups.

Motivation appeared to be unrelated to task performance, as shown in the negative and insignificant correlations with the dependent variables. This seemed to be an important finding for two reasons. First, the amount of time spent on the drawing and tangram was not related to a participant's desire to continue. Second, the quality of the work the subjects produced was not a factor that could be attributed to the desire to continue. The subjects that possessed high artistic ability were not always willing to continue the task. The type of evaluation strategy used seemed to be a factor in students wanting to continue with the tasks and return to similar tasks at a later time (Gerhert, 1986).

Bastiaens, Gulikers, Kester, and Kirschner (2006) also were concerned with student perceptions and student learning as a result of different assessment methods. They questioned how student's perceptions of the authenticity of an assignment influenced the study approach and learning outcome. They also questioned the impact of the perception of relevant assessment to the assignment on study approach and learning outcome. They were not looking to prove relationships but rather, to define the relationship and determine to what extent. The researchers noted that one important change in assessment was that they were now more

contextualized and authentic, focusing on using skills in a certain context. Assessment practices have shifted from standardized testing such as multiple choice or short response to assessments like portfolio or performance assessment. Construct validity can be defined as the assessment measures what it should measure, and consequential validity can be defined as the impact assessment has on student learning. They examined a group of 118 high school senior class students (n=118) studying social work at a vocational and educational training institute. The students were an average age of 18. The students were also in their last year of schooling and had been previously studying in a learning environment that used authentic assessment. For this study, authentic assessment was divided into four facets: 1) task, the assessment assignment that defines the content of the assessment, 2) physical context, the environment in which students have to perform the assessment task, 3) social construct, the possibilities and/or constraints of interaction during the assessment, 4) forum, the assessment method independent of the content and 5) criteria, the characteristics of the performance (product/process) that are valued. This argues that authentic assessment is a multidimensional construct and that assessment can be made more authentic in different ways by varying one or more of the facets in the assessment assignment.

The assessment for students consisted of writing a letter of application for a job and taking part in a job interview based on those letters. Both these activities were simulated in the school with the teacher taking on the role of the interviewer. Students received the assessment criteria a week before their letter was due. The students worked in groups on their professional problems. A questionnaire was also developed to examine whether and how students saw the authenticity of the task, the context, the form, and the criteria. Another questionnaire was done examining whether students perceived the instruction to have the same meaning and relevancy as the assessment. This was done with regard to what learning is valued. A revised

questionnaire was developed to determine two study approaches, surface study and deep study. There was also a course experience questionnaire to gain insight into how students felt about a learning activity. The quantitative learning was measured by two independent assessors that scores student performance based on rubrics. Data was collected for 77 of the 118 students (ndata=77).

There was a positive correlation between perception, deep study approach and the learning outcomes (task r=0.23, physical context r=0.20, form r=0.12). When the students saw the task as more authentic, they reported more use of a deep study technique. There was also a positive correlation between surface study approach and generic skill development (r=0.33, $p<0.01$) meaning that surface study approaches were used when the learning goal was a generic skill. Perhaps this is evidence of the saying "practice makes perfect." Students that used a deep study approach were better prepared for the assessment compared to students that only used a surface study approach.

Student engagement can be viewed also from the perspective of their extrinsic or intrinsic motivation. According to Amabile (1979), when individuals participate in an activity for their own sake, they are more likely to produce creative work. On the other hand, if individuals are engaged in an activity as a means to achieve some extrinsic goal, they are less creative. Amabile (1979) believed that the external or internal motivation has a great impact on student learning and student art products and process. In her quantitative study, she examined whether extrinsic motivation in the form of a final evaluation showed a decrease in creativity, unless they are told specifically how to perform creatively. Ninety-five women (n=95) enrolled in an introductory psychology course and Stanford University signed up for her experiment titled "Effects of Various Activities on Mood." This experiment was considered to be partial fulfillment of a course requirement. The experiment was also run by one female researcher. Amabile

(1979) noted that males were not used in the study to reduce outside sources of variability.

The pre-test conducted was to determine the adequacy of an art activity for this experiment. Subjects made collages from colored paper and glue that were rated on several artistic facets by artist judges. The inter-reliability was 0.77 for creating ratings and 0.72 for technical ratings. It was determined that the collage making activity was simply manipulating papers and glue to create a work of art. This allowed for the separation of technical judgments and aesthetic appeal from creative judgments. The researcher then explained to the subjects that the exercise was a pre-test for an experiment to be done the next school quarter. Each subject was to randomly choose one of five art activities and would then complete a mood questionnaire. Subjects had 15 minutes to complete their individual art activity.

There were five experimental groups: evaluation with no focus, evaluation with technical focus, evaluation with creativity focus, evaluation with specific technical focus, and evaluation with specific creativity focus. The works that fell into the specific technical focus groups were judged on the bases of six elements: the neatness of the design, the balance of the design, the amount of planning evident, level of organization in the design, presence of recognizable figures or objects in the design, and the degree to which the design expressed something to the judge. Works that fell into the specific creativity group were judged on seven elements: the novelty of the idea, the novelty shown in the use of the materials, the amount of variation in the shapes used, how asymmetrical the design was, the amount of detail in the design, the complexity of the design, and the amount of effort evident.

Fifteen judges, nine males and six females, each with at least five years of artist experience served as the judges for the experiment. Before judging, the researcher gave the judges a brief summary of the purposes and instructions given to the subjects. Each judge had to examine each of the 95 art works on 16 different categories: expression of meaning, degree of

representationalism, silliness, detail, degree of symmetry, planning evident, novelty of the idea, balance, novelty in the use of the materials, variation of shapes, effort, complexity, neatness, overall organization, creativity , and technical goodness. Before judging any of the artworks on these categories, the judge was given a brief definition of that category. Inter-judge reliability (Spearman-Brown) was calculated with the average reliability at 0.84. This is very significant considering that each judge had to rate each piece of art work alone in the room for three hours, making 1520 different judgments, spending roughly five seconds on each art work.

Amabile (1979) hypothesized that non-evaluation subjects, with the exception of the specific creativity group, were judged more creative than evaluation subjects. To test this hypothesis, a planned contrast was performed on seven different creativity dimensions: novelty of material use, novelty of idea, effort, variation of shapes, detail, and complexity. The contrast was significant, $F(1,84) = 45.81$, $p<0.001$. The researcher then continued to compare control groups and the experimental groups. When the subjects were given specific instructions to make a design, they were judged more creative than the non-evaluative subjects. “The mean rated creativity for this specific instructions group is significantly higher than that of the relevant control, $t(14) = - 3.88$, $p <0.01$; indeed, this group is higher than any other on judged creativity. In all other cases, the non-evaluation groups are significantly higher on judged creativity than the comparable evaluation groups” (Amabile, 1979, 228). The research stated that the no focus groups $t(14) = 9.44$, $p<0.001$, the technical focus groups of non-evaluation-technical focus vs. evaluation technical focus $t(14) = 2.07$, $p<0.60$, and the non-evaluation technical focus vs. the evaluation specific technical focus $t(14) = 3.62$, $p<0.01$, and the creativie focus groups $t(14) = 3.79$, $p<0.01$.

When subjects were given specific instructions on how to make a creative design, they did produce artworks that were judged as more creative than those of the non-evaluation

subjects. In the other cases, the non-evaluation groups were significantly higher on judged creativity when compared to the evaluation groups. These groups can be compared to students in the classroom that are often given art projects that will be graded, yet are left wondering the criteria for their evaluation.

Meisels, DiPrima Bickel, Nicholson, Xue, and Atkins-Burnett (2001) examined the validity of teacher judgments on a curriculum embedded performance assessment of kindergarten to third grade students in their quantitative study. Specifically, the researchers asked if teachers were able to discriminate student assessment work without the biases that could influence student outcomes. Researchers argued that teachers were capable of being valid assessors of a student's emotional, intellectual, socioemotional, and behavioral accomplishments because they observe and interact with students on a daily basis (Meisels, DiPrima Bickel, Nicholson, Xue, and Atkins-Burnett, 2001).

Seventeen teachers (n=17) in the WSS school system volunteered to participate in this study. The criteria for volunteering was that each teacher had at least two years of experience using WSS, was in the highest quarter percentile of teacher participants, and successfully completed the WSS materials for 1996-97. The teachers were chosen based on this criteria by a board. The WSS (Work Sampling System) was a low stakes, curriculum embedded performance assessment. Its primary goal is instructional assessment. It was not designed to rank students. It was designed to examine the impact of instruction. Originally, the WSS was limited to research on 100 kindergarten students. This research showed moderate to high inter-reliability rates ($r = 0.73$, $p<0.001$). The WSS included three forms of documentation – checklists, portfolios, and summary reports. The checklists for each grade consisted of specific classroom activities and learner centered expectations that were developed from state and national standards. These items measured seven domains of student development: personal and social, language and

literacy, mathematical thinking, scientific thinking, social studies, the arts, and physical development. This study only examined the language and literacy and mathematics items. These two items were the greatest interest to policy makers and school districts at the time.

Portfolios showed the efforts, progress, and achievements of a student. There are two examples of work included, core items and individual items. These are classroom produced examples of how a student functions in specific areas of learning throughout the year. They illustrate the qualitative differences in a student's work and even enable student to evaluate their own work.

The summary reports are the replacement for the conventional report card. These reports inform parents, teachers, and administrators of student progress. The summary reports are based on information recorded in the checklists, the work in the portfolio, and teacher judgments about the student.

Teachers rated a student's performance on each item of the WSS three times per year on a one to three scale with one being not yet, two being in progress and three being proficient.

The Woodcock-Johnson Psychoeducational Battery Revised (WJ-R) is an achievement test that was normed on a population of 6359 individuals chosen at random. The WJ-R scores represented standard scored that are computed by software provided by the test company. This was a method of assessment completely different from the WSS. The WJ-R was chosen as a comparable assessment to the WSS because there is no other comparable performance assessment and the WJ-R was more sensitive than other administered standard tests.

The 17 teachers implemented the WSS for the school year and continuously collected materials for the checklists, portfolios, and summer reports. The WJ-R was administered in October/November and again in April/May. Examiners were blind to the study.

Correlations were determined by comparing students' standard scores on the subjects

of the WJ-R and the WSS checklists and summary report ratings of student achievement to the corresponding WSS domains. Researchers were looking for correlations from 0.70 to 0.75. This would indicate a substantial overlap between the two assessments. It also suggests that both types of assessment contribute to students learning. If the correlations were high (>0.80), it would also suggest that the WSS did not justify its use. A low correlation (>0.30) would suggest very little overlap between the assessments and raise the question of what the predictor actually measured. Sample test sizes ranged from 75 to 94 students. Over three fourths of the correlations between the WJ-R and the WSS were in the range of 0.50 and 0.75. Forty-eight of the 52 correlations were in a moderate to high range. Correlations between the WJ-R scores in literacy and the WSS scores ranged from 0.50 to 0.80, $p<0.001$. The correlations between the WJ-R scores in broad mathematics and the WSS ranged from 0.41 to 0.83, $p<0.001$. This is strong evidence that the WSS was a valid form of assessment.

Most of the correlations in this study were reported moderate to strong. However, a few of the correlations in all of the grades was <0.50. The researchers explained this by viewing the contrast between the limited content of the WJ-R literacy items and the full range of literacy skills evaluated by the WSS. Class differences may have also affected the validity at the younger grade levels. The correlations showed a tendency to increase as the students transitioned to the next grade level.

However, there were areas of missing data due to families moving and changing residency, incomplete WSS records, and examiner variability in the administration of the WJ-R. All these different things could have affected the results of the study. The correlation attempts to demonstrate an overlap with a standardized measure of student achievement. This correlation is also similar to correlations between the WJ-R and other standardized tests. These test correlations are not provided in their entirety however, making it difficult to determine a

conclusion.

Watt (2005) conducted a quantitative study to examine teacher attitudes toward alternative assessment methods in secondary mathematics classes in Sydney, Australia. Specifically, the research questions related to teacher use of alternative assessment in math, attitudes about using alternative assessment methods, and perceived road blocks in using alternative assessment. The alternative assessment methods targeted in this study were oral tasks, practical tasks, teacher observations, student journals, student self assessment, and involving parents in the assessment process.

The mathematics staff of eleven different Sydney metropolitan schools were invited to participate in this study. Of the 11, eight were government school and three were private schools. These schools all represented a mix of government, private independent, coeducational, and single gender schools. A survey was developed for the study that consisted of eight questions, three quantitative and five open ended.

For quantitative responses to the first three questions, descriptions summarized the data based on frequency. For the five open ended questions, the answers were grouped into themes that emerged. Teachers were relatively satisfied with traditional assessment methods as a way to assess student ability. Teachers with less teaching experience appeared less satisfied with traditional test methods. Among least experienced teachers, the most common alternative assessment was observation (71%) followed closely by oral (64%) and practical (64%) tasks. Oral tasks were the most common alternative assessment used by more experienced teachers (83% for 10 to 19 years experience, 77% for 20 or more years), followed by observation (79% for 10 to 19 years, 64% for 20 or more), and practical tasks (67% 0 to 19 years, 64% 20 or more). In general, oral, practical and observation tasks were used by more teachers with student journals, self assessment. Parental assessment not frequently used. The themes that emerged concerning

why teachers would or would not use alternative assessment in the classroom fell into six categories: insufficient time for implementation, unstructured nature, unsuitable, unreliable/subjective, insufficient resources at hand, and suitable and beneficial as a reason to use alternative assessment. As we can see, there are more reasons not to use alternative assessments that there are times when it is appropriate.

These themes are all issues that will arise in implementing an alternative assessment in the classroom. The qualitative data in this study showed the reality of the current state of alternative assessment. And it further drove home the point that alternative assessment requires a lot of time, effort, resources, and forethought on the part of the instructor. The teachers for this study also provided suggestion for adapting curriculum that relies on traditional assessment methods to also incorporate more alternative assessment methods. Those suggestions were: more group work, use of observation, more practical work, develop a national curriculum, make the current curriculum less exam driven, develop assessment tasks that suit the syllabus, and change teaching methods. They also listed several alternative assessments worth more investigations such as problem solving questions, cross curricular assignments, oral projects, novel experiences, interviews, research project, and application type questions.

Morgan and Watson (2002) conducted two qualitative experiments to examine teachers' informal classroom assessment practices and the teachers' interpretation and evaluation of students' formal written math texts. Morgan and Watson (2002) were really concerned with the issue of equity in assessment. Generally, the issue of equity in assessment is only seen from the view of ensuring that all students have equal opportunity to display their progress and that these assessments do not have any bias against a particular group of students.

For student A, two teachers were getting to know students in their first year of secondary education. The teachers were both over ten years experienced as teachers and kept

their teaching practices up to date. They were fully involved in the study and knew the purpose of the study. These two teachers selected a small number of students in their classes to observe. These students were blind to the study. The researcher observed once a week and took detailed notes. Student work was photocopied and kept, and student behaviors and actions were noted.

The researcher and the teacher did peer checks with each other in formal meetings to share data and ideas. It was noted that the teacher and the researcher had the same notes and evidence. However, with the observations, it was noted that sometimes the researcher knew more about what the students had achieved but not written down, or what was before their written work. The teachers in the study were hesitant to use written work as a major source of evidence. They believed that students' mathematical thinking may not have been easily expressed in writing.

However, the researcher and the teachers did not always perceive student achievement or learning the same, as in the case of Sandra. The teacher believed that Sandra was weak in her ability to think mathematically while learning and applying new skills or grappling with new ideas. The researcher on the other hand, believed that Sandra was relatively good at math and observed Sandra using mathematical thinking skills to tackle problems on several occasions. One example was observed when she was using pentimino jigsaw pieces to create a rectangle. She used a counting squares strategy to formulate an answer. The teacher attempted to show her a different way of solving the problem and then claimed that she was not able to see new ways of problem solving (Morgan and Watson, 2002).

This case illustrates several important features: the strong influence of impressions, a teacher tendency to stick with those impressions, and the influence of positive or negative behavior.

The second study was set in the context of the standardized GCSE exam for secondary

students in England ages sixteen and older. The public education reform issued coursework, an assignment that was to be completed in class and at home and then assessed by one's teacher and external examiners. Eleven teachers (n=11) from five secondary schools each read and evaluated the texts for one of the tasks themselves and then considered what it would look like to assess it. The teachers were experienced in the general criteria for the task and the criteria were available to them if they needed it. Although the teachers appeared uncertain about this method, they eventually graded the student work with confidence and even appeared to be engaged in the assessment. The teacher interviews explored the teacher's assessment practices and identified the features in the text that teachers noted and valued. These features would also come into play as teachers formed judgments about the students and their work.

The researchers discovered diversity in the meanings and evaluations that different teachers constructed from the same texts – even with a standardized set of criteria. There were a small number of teachers that participated and a small number of student writing. The researchers were not able to quantify the differences in the grades since the grades were consistent for some texts and different for others.

There is the issue of critical thinking as well. Most of the major differences between teachers occurred when a student went outside of the norm to solve a problem. The thought process was different from the linear example the teachers were expecting even though the product was correct.

Researchers noted that assessment, even if it is based on a standard criteria, is sill dependent on interpreting student actions, some of which are significant and some which are not (Morgan and Watson, 2001). Avoiding teacher judgment does not seem possible but it is possible to be aware of our judgments. The best combatant for an inequitable assessment is the clear specification of assessment criteria and training in the use of such criteria. This can

improve the reliability of assessments.

Summary

Chapter Three was a review of the relevant research on alternative assessment. The results of these studies were analyzed and summarized based on the conclusions of those studies. This research was reviewed to determine if alternative assessments are valid and reliable. The research in the Portfolio Assessment section indicated that portfolio assessment can be a valid, reliable method of assessing student growth and development. This method of assessment also needs proper teacher instruction and plenty of practice both in working with students and evaluating student work. The Self and Peer Assessment section examined different models of self and peer assessment for reliability. The research of this section has shown that both self and peer assessments can be valid and reliable means of assessment. The Performance Assessment section reviewed performance assessment models that employed rubrics. This section also indicated that performance assessments can be reliable given proper teacher instruction and scaffolding. The last section, Effects of Assessment on Students and Teachers indicated that these different forms of alternative assessments need proper instruction and teacher awareness to be both valid and reliable. It is up to the teacher to ensure that students understand the processes of evaluation in order to receive valid student work and valid assessment results. Chapter Four provides a summary of these findings in relation to Portfolio Assessment, Self and Peer Assessment, Performance Assessment, and the Effects of Assessment on Students and Teachers. Chapter Four then will consider implications for teaching and suggestions for future research.

CHAPTER FOUR: CONCLUSION

Introduction

There are just as many ways to assess the visual arts as there are ways to create a beautiful, successful work of art. The assessment strategies a teacher chooses are dependent on who she is, what is important to her, her background, the culture that she brings into the classroom, and the students she will receive in her classroom. At the same time, assessment methods will be further defined by the national, state, and local standards; the administration; the outside community; and different cultural considerations.

Chapter One examined two different means of assessing the visual arts. On one side, educators have used standardized assessments in evaluating students in multiple content areas, including the visual arts. Other educators believed in using alternative forms of assessment to document student growth and understand student learning. Chapter one also introduced the debate between assessing the visual arts and not assessing the visual arts. On one side, teachers have stated that visual art students should not be assessed. These educators have voiced their concerns about student creativity and the subjective nature of art. Other teachers, and even students alike have voiced that the visual arts should be evaluated to gain insight into student learning and help students on their path in an art medium. Given this, Chapter One also presented the rationale for assessment in the visual art classroom and introduced the guiding question for this paper: can alternative assessments be both valid and reliable? Chapter Two explained the historical developments in the field of art education and the corresponding emphasis on assessment in education. In reviewing the history, it was noted that art education has transformed from its original stance as a service industry preparation to a means of communication and reasoning for students to a content area that has taken a backseat to other content areas in the school system. It was also noted that assessment in the visual arts has also

transformed due to new state standards. With the focus on standardized assessment in all content areas in recent years, the visual arts was also included, forcing students to participate in criterion referenced standardized assessment in several states. Chapter Three reviewed the research of alternative visual art assessment and its effects on both teacher and student. The research reviewed was organized into four sections: Portfolio Assessment, Self and Peer Assessment, Performance Assessment, and Effects of Assessment on Students and Teachers. Each of these studies was reviewed, analyzed and summarized based on the results given. The research was reviewed to examine if alternative assessment can be a valid and reliable form of assessment. Chapter Four is the final chapter of this paper. This chapter concludes the paper by revisiting the question – can alternative assessment be a valid and reliable form of assessment in the visual arts – and answering that question by using the findings from Chapter Three. This chapter also provides answers to the question, implications for teaching, and suggestions for future research.

Summary of Findings

Can alternative forms of assessment be both valid and reliable in a visual arts classroom was the guiding question for this paper. This was an important question in this time of focus and emphasis on standards and their corresponding assessment methods in the classroom. Assessment methods have evolved to become more standardized in education for several reasons: the No Child Left Behind Act, a new focus on math and science education, as well as the need for human capital in American society. However, it has not been shown that standardized assessment is the best method of assessing student knowledge and learning. According to Kennedy (1995), assessment should have different qualities to be successful: straightforward, student centered, based on coursework, differentiated, flexible, and utilizing evidence and

judgments as well as student self assessment. This differs greatly from what the nation has come to use for standardized assessment. In the classroom, teachers should be mindful of these qualities when designing meaningful formative and summative assessments. If the teacher creates assessments with these qualities, the assessments should better evaluate student knowledge, learning, and understanding. The teacher can then use this assessment information to better inform her practice and instruction.

Dorn (2003) began the Portfolio Assessment; his study examined if art portfolios could be graded and if teacher had the proper training to assess portfolios evenly and reliably. This is an appropriate start given that many art classrooms utilize the use of portfolio assessment. Beattie (1997) states that portfolios are a widely used means of alternative assessment in which students are involved in their assessment, they use tasks that represent meaningful learning activities and opportunities, and they are participating in a real world application. Furthermore, Beattie (1997) explained that portfolio assessment is not evaluated by a machine; but rather, by a human - the classroom teacher. Teachers must be able to evaluate portfolios evenly and reliably because of this The teachers in Dorn's (2003) study participated in multiple training sessions to develop assessment models, learn how to organize evidence, develop curriculum and standards, and evaluate student work. Given this rigorous training, teachers were able to grade portfolios evenly and reliably across different students and school districts. The study began the Portfolio Assessment because it demonstrated that portfolio assessment can be a valid and reliable means of assessing student learning. However, this was only made possible by teacher training and teacher attitude towards assessment.

Given the findings of Dorn (2003), Shober's (1996) analysis of the effect of teacher attitude on portfolio assessment was examined to determine how teacher attitude affects the success of a portfolio assessment. She found that teacher attitudes toward portfolio assessment

were positive and that portfolios could be used as a reliable means of tracking student growth and progress. Her study was weakened, however, when she did not publish any supportive data or her observations of student growth in the classroom. More research needs to be done to determine if teacher support of portfolio assessment lead to more reliable scoring and student improvement.

Blaikie, Schonau, and Steers (2004); Blaikie (2008), Pereira (2005) and Uram (1995) all analyzed student role in the portfolio process. Uram (1995) in particular, examined if students could evaluate their own work in a portfolio. This study also examined teacher and student attitudes toward portfolio assessment. Unlike Shober's (1996) study, the teachers in Uram's (1995) study felt that art assessment was a waste of time. This was reflected in the students' attitudes as well. Students felt unconfident in their abilities and teachers were not interested in creating dynamic assessments that could function with the curriculum. In Uram's (1995) study, the portfolio assessment was created and implemented with the design of the curriculum. After implementing a new curriculum and assessment method, along with proper teacher training, students felt more confident in their abilities. This demonstrated that portfolio assessment with proper curriculum design could increase student confidence, ability, and motivation. More research needs to be done to prove this theory, however, as this study is not generalizable based on Uram's (1995) small sample size. Uram (1995) did touch on student attitudes toward portfolio assessment. Blaikie et al. (2004), Blaikie (2008), and Pereira (2005) also investigated student opinions of portfolio assessment. Pereira's (2005) participants felt that they gained motivation and creativity through the use of portfolios and self evaluation. Students saw this form of assessment as authentic, valid, and reliable.

Blaikie et al. (2004) discovered that students believed that their portfolio assessment as well as their art classes should be solid foundations for future education. They also stated that

they felt that a portfolio enabled them to understand their progress and growth throughout the year. It was also noted that teacher should be educated and understand the content, the curriculum, and the criteria for assessment. These three items should be made explicit to the student as well. However, not all students had the same experience with the portfolio assessment. In a later study, Blaikie (2008) further questioned what high school art class and art assessment was really like for one student. This case study revealed that assessment means nothing if it is not rooted in the curriculum. Given the findings of these qualitative studies, it can be suggested that portfolio assessment is a valid and reliable means of assessment according to students; however, this assessment requires proper training, preparation and a connection to the art curriculum.

An analysis of Self and Peer Assessment was next. This section was necessary as teachers and students alluded to using self evaluation in the portfolio assessment process. As a means of monitoring student growth and progress in a portfolio, students can and did participate in self and peer assessment. Self and peer assessment are sometimes so closely tied in curriculum that they were examined together in this paper.

Several studies investigated the effects of self and peer assessment on student achievement and success. These researchers asked if these forms of assessment really could improve student performance in the classroom. Brookhart (2005) began this section by questioning if students could improve their mathematical literacy development by self monitoring and evaluating their progress. The results of this study are difficult to accept given the very small sample size and change of history of the study. Students were allowed to change the self evaluation format and the reflection questions. However, Brookhart (2005) was not the only researcher to investigate student achievement. Schunk (1996), Hassmen and Hunt (2004), and Brantmeier (2006) all examined using self assessment as an adequate predictor and

indicator of student achievement and learning. Schunk (1996) examined if self assessment could improve student learning outcomes and increate student engagement. Schunk (1996) completed two different studies. The first study revealed that self assessment increased student persistence and motivation to continue class work. The second study revealed that self assessment can also lead to enhanced self-efficacy, motivation, and performance. Schunk's (1996) findings were supported by Hassmen and Hunt (2004). Hassmen and Hunt (2004) questioned if self assessment on a multiple choice exam could further explain student understanding and learning. The findings suggested that self assessment can greatly affect different students' learning and can be an indicator of student learning style. The researchers noted that there different learning styles should be taken into account when creating assessments for the classroom.

In the classroom, self and peer assessment tend to be tightly linked. Teachers can use both peer and self assessment to evaluate student work and provide an opportunity to monitor growth and performance. These opportunities rely on student perceptions and attitudes toward self and peer assessment. Keaten and Richardson (1993) examined student perceptions and attitudes toward peer assessment. Their study revealed that students generally believed peer assessment to be fair, accurate, easy, and satisfying. In a later study, Tanner and Jones (1994) examined student attitudes, focusing on how much help students needed to make progress in the classroom. The results indicated that formative assessments help students gain access to the classroom and discover new ways of thinking that are acceptable for the curriculum. The self assessment became a change for student and teacher communication, socialization, and negotiation. This form of assessment could then give the teacher the necessary insight to help a student's development and growth. The findings of Keaten and Richardson (1993) and Tanner and Jones (1994) directly contrast Omelicheva's (2005) findings. Omelicheva (2005) examined

the effect of motivation in relation to student reliability in self and peer assessments. The results showed that self and peer assessment can both be reliable forms of assessment when evaluating a task. Students also voiced their opinions on peer assessment. These opinions differ from the opinions of participants in Keaten and Richardson's (1993) study. Omelicheva's (2005) study participants noted that peer assessment should be practiced for all students to be comfortable with the idea of peer assessment. Students stated that they feared that student bias, misunderstanding, and personal emotion could distort any reliable evaluation. These concerns should not be ignored.

Self and peer assessment can also be used to evaluate performance tasks in the classroom. Hafner and Hafner (2003) examined peer assessment as a tool for evaluating student performance. The researchers found that peer assessment was a reliable form of assessment when compared to the instructor's own evaluations. This demonstrated validity and reliability in this process of peer assessment. This study used a rubric as a checklist for both the students and the instructor. This rubric proved to be a great asset for peer assessment. Students also voiced their thoughts on the rubric, stating that the rubric was a helpful study tool that outlined the criteria and standards of the assessment. Baxter, Shavelson, Goldman and Pine (1992) assessed the scoring system for a performance assessment for a hands-on science experiment. The researchers found that a performance task can be scored reliably by different judges. This theme is seen throughout the research on performance assessment. It has been shown that performance assessment can be evaluated reliably by outside judges, instructors, peers, and students. Fuchs, Fuchs, Karns, Hamlett, and Katzroff (1999) investigated the effects of classroom based performance assessment on instruction and student problem solving. Results indicated that the classroom based performance assessment instruction increased the teacher's knowledge of what a performance assessment is and its purpose. It can also move students

toward more problem solving thinking skills as opposed to recall and basic computation. It should be noted, however, that teachers can also coach students in a performance based assessment, much like teachers can "teach to the test."

Three studies in the Performance Assessment section directly addressed the question of reliability of a performance assessment. These three studies combined suggest that performance assessment is a reliable method of assessing student learning and understanding. However, there were also areas for further research and study. Bergee (1997) investigated the reliability of music performance assessments when students were evaluated using instructor assessment, peer assessment and self assessment. The judges' were found reliable when compared with one another. However, self assessment was found to be poorly correlated with both instructor and peer evaluations. This suggests that students may not be as prepared to evaluate their own work. Yet, both self assessment and peer assessment have been seen to be beneficial assessment methods in the Self and Peer Assessment section of Chapter Three. Bergee(1997) hypothesized that students need to experience completing regular forms of assessment to gain understanding of how to do the assessment. Essentially, good practice makes perfect. Richard, Godbout, Tousignant, and Grehaigne (1999) and Nadeau, Richard, and Godbout (2008) examined peer assessment on a performance task of a team sport. Richard et al. (1999) investigated peer performance assessment in a basketball setting while Nadeau et al. (2008) examined peer assessment of a hockey practice game. Richard et al. (1999) suggested that peer assessment on a performance task is a reliable method of assessing performance. This finding was further supported by Nadeau et al. (2008). Researchers also noted in both studies that peer assessment on a performance task gave instant, objective feedback to the students and also forced the assessor to examine their own performance to invent ways to improve their learning and skills.

Students and teachers can also affect these different types of assessment; and the assessment method can have an effect on the students and teachers. Gerbert (1986) questioned the effect of evaluation on student motivation and learning in an early study. Motivation appeared to be unrelated to task performance, as shown in the negative and insignificant correlations with the dependent variables. This seemed to be an important finding for two reasons. First, the amount of time spent on the drawing and creating tangrams was not related to the desire to continue the task. Second, the quality of the work the subjects produced was not a factor that could be attributed to the desire to continue. The subjects that possessed high artistic ability were not always willing to continue the task. The type of evaluation strategy used seemed to be a factor in students wanting to continue with the tasks and return to similar tasks at a later time (Gerbert, 1986). This directly opposes the ideas of the previous studies of Omelicheva (2005) and Schunk (1996), both studies demonstrated that student motivation and engagement was increased with alternative assessment. Dochy, Gielen, Janssens, Schelfhout, and Struyven (2006) investigated the effects of end of course assessment on student performance by comparing multiple choice testing, peer assessment, case based assessment, and portfolio assessment. The findings suggest that assessments may not produce an overall effect on student learning because not all assessments in this study obtained comparable results. There were more findings on the inconclusiveness of the study than actual results. Bastiens, Gulikers, Kester, and Kirschner (2006) were also concerned with student perceptions and student learning as a result of different assessment methods. They questioned how students' perceptions of the authenticity of an assignment influenced the study approach and learning outcome. Students that used a deep study approach were better prepared for the assessment compared to students that only used a surface study approach. And students that consistently practiced a skill or technique also did well on the assessment. This could indicate

that student involvement and teacher instruction have a great effect on student learning and performance on assessments. Student engagement was also viewed from the perspective of their extrinsic or intrinsic motivation. According to Amabile (1979), when individuals participate in an activity for their own sake, they are more than likely to produce creative work. On the other hand, if individuals are engaged in an activity as a means to achieve some extrinsic goal, they are less creative. Amabile (1979) was correct. The results of her study indicated that students that were given specific instruction on a design task were rated more creative while students that were not given instruction were not. When subjects were given specific instructions on how to make a creative design, they did produce artworks that were judged as more creative that those of the non-evaluation subjects. In the other cases, the non-evaluation groups were significantly higher on judged creativity when compared to the evaluation groups. Criteria appears to be critical to the assessment process.

Meisels, DiPrima Bickel, Nicholson, Xue, and Atkins-Burnett (2001); Morgan and Watson (2002); and Watt (2005) focused on teacher attitudes and teacher judgments in their separate studies. This is very important as almost all previous research has stated that teacher instruction, preparation, and perception is crucial to the success of an assessment method. Their results showed that teachers that were involved with their students, were prepared for class, well experienced in their chosen assessment methods, and well trained in their assessment methods were more reliable in their scoring and more reliable in their evaluation of student progress and growth. However, there were moments that teachers were not able to address all student needs or notice all different areas of growth. Perhaps this is a skill that comes with time and practice; perhaps the curriculum could be revised to include all students in the learning process.

Classroom Implications

There are many negative connections between assessment and student learning in the visual arts. Amabile (1979) pointed these out: students tend to be assessed on content and matters that are easy to assess, assessment encourages students to focus on those topics that are being assessed at the expense of ignoring those that are not, the nature of assessment tasks influence the approach to learning that students adopt, students who perform well on tests and examinations retain deep misconceptions about key concepts in the content areas they have passed, students give precedence to assessment which counts toward their final assessment, and successful students look for clues and cues from teachers that enable them to identify what is important for assessment purposes. There is much that stands in the way of growth through appropriate assessment. Based on the research findings of Chapter Three, there are several ways to conquer the negative side of assessment. One way is to incorporate alternative assessment into the visual art curriculum. Assessing a student's art work is a complex, potentially confusing, and vital to the development of student creativity, learning, and motivation. The researchers used a variety of assessment methods, focusing mainly on portfolio assessment, self assessment, peer assessment, and performance assessment to engage the subjects and monitor student learning, motivation, and attitude in regards to the validity, reliability and purpose of the assessment.

Teachers should select reasons for using a portfolio assessment based on student needs and educational goals. Portfolios can be a useful additional or alternative form of assessment in art. According to Uram (1995) students become more responsible for their work as they take ownership of their learning by setting goals and monitoring their progress. Through the use of the portfolio over a length of time, students are allowed to preserve their work, reflect upon it, and set new goals for themselves based on their personal strengths, weaknesses and needs.

Attention should also be given to the art curriculum. The curriculum should be designed to use the portfolio assessment as a means to store, review, record, and preserve many opportunities, versions and revisions of student artwork. This can then incorporate higher order thinking skills and develop metacognitive skills (Uram, 1995). Uram (1995) recommended that teachers decide on eight questions: 1) the purpose of the portfolio, 2) who will contribute to the portfolio (student, teacher, parent), 3) how one will establish criteria for portfolio development, 4) what students will include in their portfolios, 5) how students will review and add to their portfolios, 6) how one will engage students in self evaluation, 7) how one will evaluate the portfolio (what criteria to use), and 8) how one will use the portfolios to set future goals.

Aschbacher (1992) goes so far to say that the assessment in a portfolio exists only when an assessment purpose is identified, criteria of methods for determining what is put into the portfolio, by whom, and when, are explained, and criteria for evaluating the collection or individual pieces or work are identified. Without these, there is no assessment. This assessment method involves much detail and planning on the part of the teacher. Teachers need to make provisions and be proficient in their evaluations and their instruction with the students.

In regards to self and peer assessment, it should be noted that these assessment methods should be analyzed and scrutinized to ensure that these methods are appropriate for a specific student or classroom (Schunk, 1996). Keaten and Richardson (1993) and Tanner and Jones (1994) supported this stating that teacher instruction and preparation was a key in creating a safe environment that can allow for peer assessment. By creating that safe environment, this form of assessment could then give the teacher the necessary insight to help a student with his or her development and growth. This safe environment was also critical to create a space where students could feel comfortable practicing and participating in self and peer assessment. As Omelicheva (2005) noted, students felt that student bias,

misunderstandings, and personal emotion and contempt could disturb any form of a reliable evaluation. A community of learners can be established that discusses and informs students of possible biases that can exist in the classroom including gender, race, ethnicity and socioeconomic status. Educators should also be aware that using peer assessment makes a student's personal information public and readily available to a student's fellow classmates. This could potentially be an uncomfortable situation for some student. Peer assessment needs to be practiced continually for all students to feel comfortable with the assessment method. This can then lead to more reliable results and evaluations.

Amabile (1979) and Schunk (1996) also alluded to the importance of learning goals in assessment methods. Even though Amabile (1979) and Schunk (1996) examined two different forms of assessment – self assessment and performance assessment – the findings revealed that setting learning goals enhanced the learning experience for the students. By giving students a learning goal, student self-efficacy, motivation, skill development, and task orientation was improved. These improvements were positively influenced by letting students self assess and reflect on their performance skills and growth. It is possible that by emphasizing the learning goals and objectives, students focus less on getting the correct answer and more on ways to solve problems (Schunk, 1996). It should be noted though that self and peer assessment should be supplemented and supported with proper classroom instruction so that students can see and experience the benefits of monitoring and recording their growth. If there is not proper classroom instruction and support, students can fall into a cycle in which failure leads to negative self esteem and perception, lack of motivation and more failure.

In the area of performance assessment, the most widely used tool for performance assessment in the research was the rubric. A great rubric communicates what sufficient, good, and excellent student work should look like. The rubric can also involve the student in

constructing the criteria and evaluation method. By using a rubric as part of the performance assessment process, teachers can involve the students in constructive learning and active reflection and self evaluation. According to Brookhart (1999), a rubric is a checklist that describes the criteria for good student work. This checklist should be directly related to the curriculum, the learning objectives, the critical thinking, or the goals that the teacher would like students to acquire. For an even more effective rubric, the teacher should share the rubric with the students ahead of time. This will enable students to better understand what is expected of them. This process of sharing and creating a rubric with students can also lead to enhanced student motivation, engagement, and learning. Bergee (1997);Richard et al. (1999); Hafner and Hafner (2003); and Nadeau et al. (2008) employed a rubric for observing and collecting information on student performances. While the research suggested that rubrics are a useful tool for student evaluation, the teacher participants of these studies noted that students need time and practice using this assessment method to reliability evaluation themselves and other peers. Teachers should be aware that students may have a difficult time defining the criteria or skills they are observing. Students could also potentially lose interested and motivation while teachers are training students to effectively use a rubric for performance assessment (Nadeau, Richards, and Godbout, 2008).

In the visual art classroom, teachers need to define and determine what their role is in the assessment process. Teachers are responsible for knowing their students, understanding and interpreting student learning, adapting new curriculum based on student needs, and monitoring student progress and growth. According to Grehaigne and Godbout (1998), teachers need to provide appropriate assessment for their students to ensure student success. This can be done through communicating explicit expectations, helping students manage their own learning and observing students during certain tasks. Teachers should create assessments that

are regular and ongoing, connected to daily instruction, and easily visible and evaluated by the teacher and/or the students. The research has also shown that when properly practiced and experienced, students are very capable of evaluating their own performance, development, and learning. It is up to the teacher to provide that experience and practice for their students.

Suggestions for Further Research

There are numerous problems in the world of visual art assessment that are worth more investigation. As shown in the research of Chapter Three, assessment is more than just choosing an alternative assessment and going with the flow. There are multiple areas for further research. One area may be parent attitudes and opinions on alternative assessment as well as teacher attitudes towards alternative assessment.

With the continuing push of the No Child Left Behind Act, there is also interest in the aspect of criterion referenced standardized testing in the visual arts as well as state standards in visual arts curriculum. Two areas of further research my include the effects of state standards on alternative assessment as well as the possibility of alternative assessment becoming a form of standardized assessment for certain performance based content areas. Washington State has already begun using criterion referenced testing for the visual arts (www.k12.wa.us). However, no research or evidence was found to support its use in the classroom.

The research of Chapter Three highlighted several areas for more investigation in regards to teacher training and development. Teacher training to evaluate portfolios is a must in regard to portfolio assessment to ensure rater reliability and validity of the portfolio as an assessment tool. Many students in these research studies had previously followed the pattern of creating consecutive artworks with no means of storing them or reviewing them at a later date. In implementing this type of assessment, teachers need to take notice of their pedagogy and

reasons for employing a portfolio assessment. Dorn (2003), Shober (1996), and Pereira (2005) suggested that further research should be devoted to examining teacher training methods in evaluating student portfolios as well as teacher implementation of the assessment method. Uram (1995) noted that teacher training should be examined with the curriculum of the classroom. This is further supported by the research of Blaikie, Schonau, and Steers (2004) and Blaikie (2008). Further research should be devoted not only to teacher training but also to appropriate assessment with appropriate curriculum. This was a theme seen in all the forms of assessment examined in the research. Further research should also be conducted to examine alternative assessment as an indicator of ability or academic achievement. Brantmeier (2005) and Hassmen and Hunt (1994) both suggested further study o examine instructional practices that improve alternative assessment and also whether these practices improve student achievement and performance.

Perhaps the biggest area for further research involves the role of the teacher in the implementation and completion of the assessment. Nearly every research study noted that the teacher's role in the assessment process is to fuse instruction, curriculum and assessment into a meaningful learning experience. This can be further investigated by examining different effective strategies and instructional practices teachers can employ in the classroom to create this meaningful learning experience. Furthermore, there is very little research available that directly investigates the assessment methods in the visual arts. Perhaps that is even an area of further study, simply examining and analyzing the current trends and assessment methods that are being used in visual art classrooms around the nation.

Conclusions

Chapter One examines the reasons for an investigation of the research regarding alternative assessment. It explained the beginning of standard assessment in the visual arts; it defined the differences between standardized and alternative assessment. It also introduced and defined different methods of alternative assessment and their processes. Chapter One also discussed the opposition to alternative assessment and gave the reader limitations for the research review. Chapter Two explained the development of art education in the United States, the formation of assessment in the visual arts, and the formation of standardized art assessment in the United States, particularly in Washington State. Chapter Three reviewed the research of alternative assessment. The research in Chapter Three was divided into four themes: Portfolio Assessment, Self and Peer Assessment, Performance Assessment, and The Effects of Assessment on Students and Teachers. These themes were used to answer this paper's guiding question: what alternative assessment methods are valid, reliable, and engaging. The research reviewed in the Portfolio Assessment section suggested that portfolio assessment can be a reliable, useful tool for evaluation that is unique to each student and enable students to understand their own progress and growth in the class. The research in the Self and Peer Assessment section found that self and peer assessment can be a reliable form of assessment when used properly; including student training, community building, and teacher preparation. This form of assessment can help students focus on the task at hand and their own learning. The research reviewed in the Performance Assessment section was similar to this. The research in the Performance Assessment section suggested that performance assessment can be evaluated reliably by instructors, outside judges, peers and students. However, this assessment method needs much teacher preparation, explicit instruction, and practice to become more reliable in the classroom over time. The research reviewed in the Effects of Assessment on Student and

Teachers suggested that alternative assessment methods have a positive effect on motivation, learning, and creativity. The research was reviewed to examine if alternative assessment is reliable and engaging. Chapter Four included a summary of the findings based on the four sections from Chapter Three, implications for classroom practice, and suggestions for further research.,

Beattie (1997) saw alternative means of visual art assessment as the most effective strategy in evaluating student artwork. Portfolio assessment, self and peer assessment, and performance assessment when combined with different methods of assessment appear to be an ideal for evaluating student art work. However, with the No Child Left Behind Act and the increased focus on standardized forms of assessment, the classroom culture is moving more towards standardized assessment. By using different alternative assessments, students have to do more than surface tasks and memorization. They must organize, synthesize, and clearly communicate what they understand and what they have learned. This process requires constant student reflection and self monitoring (Slater, 1996). Introducing students to alternative assessment can help to engage students, increase student motivation, and produce student creativity. The more students experienced and practiced these different alternative assessment, the more reliable the assessment became and the more confident the students became in the evaluation process. Alternative assessments can aid students in realizing their growth and their areas for improvement, their progress, and their potential.

WORKS CITED

Amabile, T.M. (1979). Effects of external evaluation on artistic creativity. *Journal of Personality and Social Psychology*. *37*(2), 221-233.

Angelo, T. A. (1995). Reassessing and defining assessment. AAHE Bulletin (Nov.), 7-9.

Aristotle. (1941) *Politics. The Basic Works of Aristotle*. Edited by Richard MeKeon. Random House. New York. 1306-1308. (Original written 384-22 BCE).

Arts K-12 Learning Standards. (n.d.). Retrieved from The Office of Superintendent of Public Instruction Official Website: http://www.k12.wa.us/Arts/Standards/default.aspx.

Aschbacher, P.R., Herman, J.L., & Winters, L. (1992). *A Practical Guide to Alternative Assessment*. The Regents of the University of California. Oakland.

Bartels, L.K., Bommer, W.H., & Rubin, R.S. (2000). Student performance: Assessment centers versus traditional classroom evaluation techniques. *Journal for Education for Business. 75*(4), 198-201.

Bastiaens, T.J., Gulikers, J.T.M., Kester, L. & Kirschner, P.A. (2006). Relations between student perceptions of assessment authenticity, study approaches, and learning outcomes. *Studies in Educational Evaluation. 32*, 381-400.

Beattie, D.K. (1997). The feasibility of standardized performance assessment in the visual arts: Lessons from the Dutch model. *Studies in Art Education. 34*(1), 6-17.

Blaikie, F., Schonau, D., Steers, J. (2004). Preparing for portfolio assessment in art and design: a study of the opinions and experiences of exiting secondary school students in Canada, England and The Netherlands. *International Journal of Art and Design Education. 23*(3), 302-315.

Blaikie, F. (2008). Emma's trouble with Mr. Robson's teaching. *Canadian Review of Art Education, Research and Issues. 35,* 19-36.

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System. 34*, 15 - 35.

Brookhart, S. (1999). The art and science of classroom assessment: The missing part of pedagogy. *ASHE-ERIC Higher Education Report. 27*(1). The George Washington University, Graduate School of Education and Human Development. Washington, DC.

Brookhart, S., Andolina, M., Zuza, M., Furman, R. (2004). Minute math: An action research study of student self-assessment. *Educational Studies in Mathematics, 57*(2), 213-227.

Casbon, C. (1989). Involving pupils in their own learning through records of achievement. *British Journal of Physical Education.* 90-92.

Cooper, B. (1994). Authentic testing in mathematics? The boundary between everyday and mathematical knowledge in National Curriculum testing in English schools. *Assessment in Education. 1*(2), 143-166.

Dewey, J. (1958) *Art as Experience*. Capricorn Books. New York. (Original work published 1934).

Dewey, J. (1954) Art and Education. Art and *Education: A collection of essays by John Dewey and others.* 3rd Ed. Barnes Foundation Press. Merion.

Dochy, F., Gielen, S., Janssens, S., Schelfhout, W., & Struyven, K. (2006). The overall effects of end-of-course assessment on student performance: A comparison between multiple choice testing, peer assessment, case-based assessment, and portfolio assessment. *Studies in Educational Evaluation. 32*, 202 -222.

Doppelt, Y. (2009). Assessing creative thinking in design based learning. *International Journal of Technology and Design Education. 19*(1), 55-65.

Dorn, C.M. (2003). Models for assessing art performance (MAAP): A K-12 Project. *Studies in Art Education. 44*(4), 350-370.

Ecker, D.W. & Eisner, E.W. (1966) Some historical developments in art education. *Concepts in Art and Education: An Anthology of Current Issues.* Edited by George Pappas. Macmillan, New York.

Ecker, D.W. & Eisner, E.W. (1966). What is art education? *Readings in Art Education*. Blaisdell Publishing Company. Waltham.

Efland, A. (1990). *A History of Art Education: Intellectual and social events in the teaching the visual arts.* Teachers College Press. New York.

Eisner, E.W. (2002). *The Arts and the Creation of Mind*. Yale University Press. New Haven.

Eisner, E.W. (1963). Evaluating children's art. Some historical developments in art education. *Concepts in Art and Education: An Anthology of Current Issues*. Edited by George Pappas. Macmillan, New York.

Gerbert, G.L. (1986). Effects of evaluative statements on artistic performance and motivation. *Studies in Art Education. 27*(2), 61-72.

Greer, L. (2001) Does changing the method of assessment of a module improve the performance of a student? *Assessment &Evaluation in Higher Education. 26*(2), 127-138.

Greer, W.D. (1984). Discipline based art education: Approaching art as a subject of study. *Studies in Art Education. 25*, 212-218.

Grehaigne, J.F. & Godbout, P. (1998) Formative assessment in team sports in a tactical approach context. *Journal of Physical Education, Recreation, and Dance. 699*, 46-51.

Grehaigne, J.F., Godbout, P., Richard, J.F., & Tousignant, M. (1999). The try-out of a team sport performance assessment procedure in elementary and junior high school physical education classes. *Journal of Teaching in Physical Education. 18*(3), 336-356.

Gruber, D. D. & Hobbs, J.A. (2002). Historical analysis of assessment in art education. *Art Education. 55*(6), 12-17.

Hafner, J.C. & Hafner, P.M. (2003) . Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education. 25*(12), 1509-1528.

Hargreaves, D.J. (1997). Student learning and assessment are inextricably linked. *European Journal of Engineering Education. 22*(4), 401-409.

Keaten, J.A. & Richardson, M.E. (1993) A field investigation of peer assessment as part of the student group grading process. Presented at the Annual Meeting of the Western States Communication Association. Albequerque, NM.

Lanier, V. (1991) *The World of Art Education According to Lanier*. National Art Education Association. Reston, VA.

Ledderman, A. (1988) Art for the real world. *Beyond dbae: The Case for Multiple Visions of Art Education*. Edited by Judith Burton, Arlene Lederman, and Peter London. University Council on Art Education.

Lidston, J. (1988). A conversation with myself. *Beyond dbae: The Case for Multiple Visions of Art Education*. Edited by Judith Burton, Arlene Lederman, and Peter London. University Council on Art Education.

London, P. (1988). To gaze again at the stars. *Beyond dbae: The Case for Multiple Visions of Art Education*. Edited by Judith Burton, Arlene Lederman, and Peter London. University Council on Art Education.

Lowenfeld, V. &Brittain, W.L. (1982). *Creative and Mental Growth*. Macmillan. New York.

Nadeau, L. (2008). The validity and reliability of a performance assessment procedure in ice hockey. *Physical Education and Sport Pedagogy. 13*(1), 65-83.

Omelicheva, M. (2005). Self and peer evaluation in undergraduate education: Structuring conditions that maximize its promises and minimize the perils. *Journal of Political Science Education. 1*(2), 191-205.

Pereira de Eca, M.T.T. (2005) Using portfolios for external assessment: An experiment in Portugal. *International Journal of Art and Design. 24*(2), 209-219.

Rush, J.C. (1987). Interlocking images: The conceptual core of a discipline-based art lesson. *Studies in Art Education. 28*(4). 206-220

Schafer, W.D., Swanson, G., Bene, N., & Newberry, G. (1999). Effects of teacher knowledge of rubrics on student achievement in four content areas. Office of Educational Research and Improvement. Paper presented at the Annual Meeting of the American Educational Research Association. Montreal, Quebec, Canada.

Shober, L. (1996). A portfolio assessment approach in narrative writing with the cooperation of a fourth grade target group. M.S. Practicum. Nova Southeastern University.

Slater, T.F. (1996). Portfolio assessment strategies for grading first-year university physics students in the USA. *Physics Education. 31*(5), 329-333.

Somervell, H. (1993). Issues in assessment, enterprise and higher education: The case for self-, peer and collaborative assessment. *Assessment & Evaluation in Higher Education. 18*(3), 221-233.

Spring, J. (2005). *The American School, 1964-2004*. McGraw-Hill. New York.

Steers, J. (1994). Art and design: Assessment and public examinations. *Journal of Art and Design Education. 13*(3), 287-298.

Tanner, H. & Jones, S. (1994). Using peer and self-assessment to develop modeling skills with students aged 11 to 16: A socio-constructivist view. *Educational Studies in Mathematics. 27*(4), 413-431.

Topping, R.J. (1990). Art education: Crisis in priorities. *Art Education. 43*(1), 20-24.

Uram, S. (1993). Art portfolios: Elementary assessment. Masters Action Research Project. Saint Xavier University. Rockford, IL.

Watt, H.M.G. (2005). Attitudes to the use of alternative assessment methods in mathematics: A study with secondary mathematics teachers in Sydney, Australia. *Educational Studies in Mathematics. 58*(1), 21-44.

ASSESSING THE VISUAL ARTS:

VALID, RELIABLE, AND ENGAGING STRATEGIES

by

Alexandria English

A Project Submitted to the Faculty of
The Evergreen State College
In Partial Fulfillment of the Requirements
For the Degree
Master in Teaching
2010

This Project for the Master in Teaching Degree

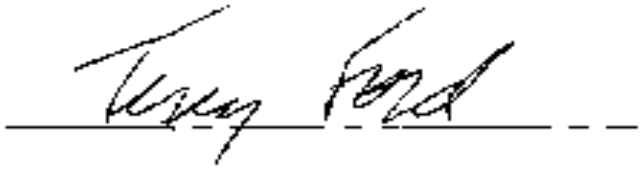
by

Alexandria English

Has been approved for

The Evergreen State College

by



Dr. Terry Ford

June 2010

ACKNOWLEDGEMENTS

I would like to thank everyone that has helped me on this long journey. An extra thanks to Dr. Terry Ford for reading pages and pages of rough drafts and revisions to get me to this point. Also, thanks to all the faculty members of the MiT program for their dedication and passion for their students.

I would especially like to thank my family for their never ending belief and support. Thank you to my fellow cohort members for keeping me focused, motivated, and sane. And of course, thank you, Christopher, for washing dishes and cleaning the house when I was too tired, being my shoulder to cry on, forcing me to have fun when I needed it, and letting me sleep in on those rare weekend mornings.

ABSTRACT

This paper attempts to answer the question what are alternative assessment methods that teachers and students find reliable and valid. To answer this question, this paper discusses the history of art education and examines research that investigated portfolio assessment, self and peer assessment, performance assessment, and teacher and student perceptions of alternative assessment.

While there is still more research to be done in the area of visual arts assessment, some conclusions can be drawn from the research investigated here. Portfolio assessment, self and peer assessment, and performance assessment were all found to be reliable methods of assessing students. However, there is more to be discovered in regard to teacher instruction practices and curriculum development. There are also implications for teachers and instructors that plan to use alternative assessment in their classrooms.

TABLE OF CONTENTS