PORTFOLIO ASSESSMENT: INDIVIDUAL CLASSROOMS AND WIDE-SCALE APPLICATIONS

by

Susan McLaughlin

A Project Submitted to the Faculty of
The Evergreen State College
In Partial Fulfillment of the Requirements
for the degree
Master in Teaching
1999
This Project for the Master in Teaching Degree

by

Susan McLaughlin

has been approved for

The Evergreen State College

by

Dr. George Freeman, Jr., Ph.D.

2/25/99
Abstract

The purpose of this paper is to explore the mechanical aspects of portfolios and examine significant questions and concerns regarding the use of them. Empirical and anecdotal evidence is provided as well as an outline of significant events that have occurred in American schools. In chapter 3, portfolios used as instructional tools within the classroom will be investigated first and then wide-scale applications will be explored. The conclusion will provide an overview of the research found and also explore different possibilities the portfolio movement may take.
# Table Of Contents

- **Acknowledgments** 5
- **Chapter 1: Introduction** 6
- **Chapter 2: Historical Perspectives on Assessment** 10
- **Chapter 3: Integrative Review of the Literature** 23
- **Chapter 4: Conclusion** 49
- **References** 53
Acknowledgements

I want to thank my husband, Michael, for his unwavering support and encouragement. I also want to acknowledge my family, friends, colleagues, and faculty advisors for their enormous amount of faith and enthusiasm. I want to give special gratitude to Dante Salvatierra for generously supplying research assistance and to Dr. George Freeman, Jr., Ph.D., for his invaluable advice, wisdom, and inspiration.
CHAPTER 1: INTRODUCTION

Today, testing is a vital issue in educational reform. There have been many debates on how to improve the quality of public schooling in our country. Currently, at least forty states are in the process of implementing or developing alternatives to traditional forms of testing (Castiglione, 1996). Performance-based or authentic assessment has gained popularity as criticisms of traditional multiple-choice type tests have increased. One of the more popular assessment alternatives is the portfolio. The use of portfolios to assess student achievement is becoming more widely used.

Many definitions have been given to describe portfolios. Hill, Kamber, and Norwick (1990) describe a portfolio as a:

. . . colorful collection of a person’s work that shows his or her thoughts, interests, efforts, and goals, in many different environments. Portfolios help learners see how they think, feel, work, and change over a period of time. (p. 118)

The contents found in a portfolio can vary depending upon the student, the teacher, or the school district. It may consist of the student’s best pieces or contain all of the work completed for the class. Sometimes portfolios include peer evaluations, self-reflections, and goals set by the student. Besides samples of the student’s work, a portfolio may also include photographs, video tapes, and journals that documents his or her understanding of a topic.

When implemented, for example, a literacy portfolio in a classroom usually consists of students’ best pieces, initial drafts, and reflections. Every month or so the classroom teacher assigns a new writing project. Throughout the year, these projects cover a variety of different writing genres. Students constantly revise and edit their work, keeping all of it in writing folders. Periodically, students take
their portfolios home to share with their parents. Entries contained in the portfolios provide valuable information about students’ strengths and weaknesses allowing teachers to make informed decisions about their students’ abilities. Teachers can then base their lessons on student need linking instruction with assessment. In a wide-scale assessment program, the portfolios could then be scored by trained judges. Just as with standardized tests, the results would be published.

Supporters of portfolios like Gilman and Hassett (1995) claim that, “Portfolios measure learning holistically, instead of isolating those bits and pieces of factual knowledge that are emphasized on skill-based achievement tests” (p. 313). Portfolios are also more authentic because they measure what is actually being taught in the classroom. On the other hand, those who support more traditional forms of testing argue that portfolios are “less efficient” and “time consuming” and therefore are no match to standardized tests especially in large-scale applications (Castiglione, 1996). Research has also shown that the evaluations in portfolios can be subjective and the results often lack reliability and validity (Koretz, Klein, McCaffrey, & Deibert, 1992).

Currently and in the past, standardized testing has been the preferred instrument used to measure student intellect. In the areas of management of instruction and public accountability, Resnick (1979) identifies the function tests serve in our schools today. Resnick identifies three types of management of instruction: sorting, grading, and monitoring. Tests used for sorting purposes are often administered before a course or program begins. The results determine the various levels students should enter at or even if they can enroll in the class at all. Aptitude and intelligence tests fall under this category. Tests given at the end of a course or program are used to determine what students
have learned. Scores obtained from these kind of assessments, often referred to as achievement tests, are used for grading purposes. The last type is essentially used to monitor students progress during a course. Curriculum reference tests assist educators in making informed decisions about what to teach next.

The second category, public accountability mainly use achievement tests, also known as norm-referenced tests, to determine how well students and schools are doing. Norm-referenced tests, for example, are used to compare individual students to others. Students' scores are compared with a "norm group" which may have taken the test up to seven years earlier (Herman, 1992). In *Taking Aim at Testing*, Rothman (1996) writes, "Teachers say they seldom find norm-referenced tests useful for diagnosing students' strengths and weaknesses" (p. 28). The scores obtained mainly rank and classify students. Critics of traditional forms of testing also argue that some students are "unable to demonstrate their learning on tests because of debilitating test anxiety and to problems of cultural, racial, or gender bias in tests" (Worthen, Borg, & White, 1993, p. 5).

As the controversy over testing continues, educators, now, more than ever, need to be informed about different ways of measuring student achievement. Recently, portfolios have gained an enormous amount of attention requiring educators to examine this instrument in more detail. "In the 1990's," states Fontana, (1995) "portfolios are not only synonymous with authentic assessment but have taken their place at the top of the list ahead of other authentic techniques" (p. 25). Authentic, known also as performance-based assessment, requires individuals to actively demonstrate their knowledge by reflecting and synthesizing what they have learned. More and more schools are turning to this
form of assessment as an alternative to standardized testing. Large-scale portfolio assessment is currently being used in both Vermont and Kentucky. In Vermont students in fourth and eighth grades produce portfolios and Kentucky requires portfolios in grades four, eight, and twelve. These portfolios are used for assessing students in both mathematics and English composition. In classrooms across the country, educators are choosing portfolios to assess their students especially in writing. Portfolios are slowly filtering into our educational system. For this reason, it is important that this topic be explored by analyzing whether portfolios are the most effective method in assessing student abilities.

As a teacher preparing to enter into the field of education, I wanted to examine portfolio assessment both in individual classroom settings and in wide-scale applications. More specifically focusing on the use of portfolios as instructional and accountability tools. The guiding question throughout this paper will be: What are the mechanical aspects of portfolios and what has research shown in terms of significant questions and concerns regarding the use of them? This project will investigate both empirical and anecdotal evidence available in the K-12 educational system. Before analyzing the research on portfolios used within the classroom and in wide-scale applications, this paper will first outline significant events in American schools. The conclusion will summarize the findings found in the research section and also explore different possibilities the portfolio movement may take.
CHAPTER 2: HISTORICAL PERSPECTIVES ON ASSESSMENT

Throughout history, different methods have been used to assess student achievement. Whether using oral questioning, standardized tests, or portfolios, the chosen assessment tool should provide accurate information about student academic abilities. In *Assessing Students Performance* Wiggins (1993) explains that the word assess comes from the Latin word *assidere* which means to "sit with." "In an assessment, one sits with the learner. It is something we do *with* and *for* the student, not something we do *to* the student" (p. 14). In this sense, educators should act more as facilitators or coaches rather than transmitting skills and knowledge. An effective assessment program should provide evaluative information about students' skills which educators can in turn use to make informed decisions about what to teach next.

Although the portfolio movement has gained a considerable amount of attention, standardized tests "have enjoyed unprecedented and probably undeserved popularity since the 1970's" (Gilman & Hassett, 1995, p. 310). During the last 70 years testing measurements have been influential in shaping our educational system (Resnick, 1982, p. 173). Precise measurements were developed by both educational reformers and psychologists to differentiate between student abilities. Today individuals of all ages are tested for various reasons. How is it that standardized testing became so popular, and, more importantly, what events and trends led us to where we are now? Examining the historical roots of assessment allows us to see how our educational system has evolved and what impact it has had on our society.

**Seventeenth Century**

Schooling in America during the seventeenth century mainly occurred in the home. The family was responsible for educating their youth. Parents and
grandparents were the children’s primary source in learning how to read and write. Religion played a major role in educating the youth. The Puritans taught their children to read the holy scriptures in hopes of maintaining religious order. The goal of Colonial education was primarily to save souls and to create model citizens. During Sunday services children were required to memorize scriptures and often attended additional religious classes (Spring, 1997).

Another means of educating children during the Colonial era was through apprenticeship programs. Children were sent away to live with a mentor who not only taught them a vocational skill but also how to read and write. In larger communities, children attended Dame schools. A local woman would teach reading and writing to children in her home. Whether it be in the home, at church, through apprenticeship programs, or a dame school, how well a child learned how to read and write depended greatly upon the skills of the teacher. As a result, there was no consistency in the quality of education the children received.

The fist text book used to educate children was the New England Primer. This instructional material was widely used until Noah Websters’s American Spelling book was published in 1783. The primer consisted of the alphabet and religious verses. Children were expected to memorize the content and not question or analyze it.

One major milestone in developing formal schooling in America occurred in Massachusetts in 1647. “The Old Deluder Saton Law” required communities of 100 households or more to pay a teacher to teach reading and writing. By the mid-seventeenth century, grammar schools were established throughout New England (Spring, 1997) Colonial education greatly impacted American public schools. It was during this time period that education of the youth switched from
being the families responsibility and moved toward being the states.

Eighteenth Century

During the eighteenth century, Thomas Jefferson and Benjamin Franklin were both instrumental in moving schools toward more nonsectarian practices and pedagogy and saw the benefit of providing education to all social classes.

Thomas Jefferson saw that Americans would benefit greatly if everyone was educated rather than just the elite. He also saw the need to provide instruction that was not just religiously based. In 1751, Benjamin Franklin’s Philadelphia Academy opened. Franklin’s academy appealed to merchants and others who wanted to learn more practical skills. Unlike the Latin grammar schools, students were able to choose from a variety of subjects instead of only focusing on Latin and Greek.

In 1783, Noah Webster’s American Spelling Book replaced the New England Primer. By 1875, 75 million copies had been sold (Spring, 1997). This text book included the alphabet, short stories, and rules for speaking. In addition, it also promoted patriotism and a national culture. Assessment continued to consist of primarily oral examinations. The influences made by Thomas Jefferson, Benjamin Franklin, and Noah Webster carried over into the early decades of the nineteenth century.

Nineteenth Century

Two of the most influential educational leaders in the mid-nineteenth century were Horace Mann and Henry Barnard. Both fought vigorously for nonsectarian, locally controlled, and tax-supported schools. Their vision consisted of public schools available to all children regardless of their academic backgrounds (Borrowman, 1973).

In Massachusetts, Horace Mann helped establish America's first State Board
of Education which there he became secretary of in 1838. Between 1838-42, Henry Barnard served the same position in Connecticut. While serving in these positions, Mann and Barnard collected data on school conditions. They noted deficiencies and also made suggestions for improvements. Their goal was that all schools would be properly maintained and equipped with qualified teachers and sufficient facilities and materials. In Mann’s first Annual Report he personally investigated 800 schools while Barnard professed to have personally researched over 200 schools in his report. Besides personally inspecting schools in their respective states, they also received reports on other schools. Combined they examined data collected from over 1700 schools.

These educational reformers reported their findings in "inquiries" or "circulars" which were distributed to the community. According to Henkin and Ignasias (1978) these reports often “promoted competition among school committees” (p. 242) since they often made comparisons between schools. Mann and Barnard frequently challenged teaching practices and the curriculum being taught. They suggested better teacher training and a more practical curriculum which would give students the skills needed to function well in society.

Prior to the late 1800’s, oral examinations were frequently used in American Schools. This was the primary assessment used until Horace Mann introduced achievement tests. Horace Mann, often referred to as the the father of the public school, was a strong advocate of using written tests over oral examinations. He argued that this new instrument was more efficient in providing objective information on what children know. Mann argued, “The tests would determine, beyond appeal for gainsaying, whether the pupils have been faithfully and completely taught” (Rothman, 1995, p. 34). In 1845, the first standardized
achievement test was administered in Boston, Massachusetts. The results indicated many failures which sparked criticisms against school officials. These findings also helped Horace Mann gain support in his vision of creating a common school. Horace Mann was instrumental in getting schools to move toward using written exams and holding schools accountable for students' academic performance (Rothman, 1995).

Not everyone agreed with Mann and Barnard's vision of establishing a system of free education. Private schools and religious leaders opposed the idea and tax payers complained that supporting public schools would be too costly. Business leaders also questioned the idea of educating everyone. They were concerned that over educating citizens would lead to discontent and would completely diminish child labor.

Although there was opposition, Mann and Barnard's persistence prevailed. Their vision appealed to people who hoped that free schools would improve the lives for all children. Public education would improve society and increase social mobility. Other states started following Mann and Barnard's lead. By 1876, a new state could not enter the union unless it provided free education (Nygren, 1978).

Another prominent figure in the history of American education was Joseph Mayer Rice. On January 7, 1892 Rice, known as "the real inventor of the comparative test," (Resnick, 1982, p. 180) published his findings. Rice interviewed teachers and parents in 36 cities and administered spelling tests to 29,000 students (Cremin, 1973). He found there wasn't a correlation between spelling ability and the amount of instructional time spent in the classroom. He referred to "... school administrators as political hacks and teachers as incompetents who blindly led innocent students in sing song drill and rote
repetition" (Worthen, et al., 1993, p. 17). As a result, the public supported Rice and began to question teachers' techniques.

He proposed that licensing for teachers become more strict and that they constantly seek improvement. He believed more guidelines and supervision would be needed to improve the quality of education in the United States. School administrators and teachers throughout America were pressured into rethinking their strategies. These influences carried over into the twentieth century.

Twentieth Century

One of the most influential events that greatly impacted the role of education in the United States during the twentieth century was the creation of the intelligence test. The Minister of Public Education in France asked Alfred Binet to create an instrument which could identify learning disabled students. The goal was to create "a more consistent, less biased means of evaluation" (Siegler, 1992, p. 182). Students identified as being learning disabled would then be removed from normal instruction and be placed in separate classes that could address their special needs. In 1905, Alfred Binet designed and published an intelligence scale which consisted of thirty tasks. This new form of assessment essentially standardized students answers in areas of memory, reasoning, and comprehension (White & Hall, 1980).

In 1916, Lewis Terman at Stanford University revised Binet's scale. It became known as the Standford-Binet Intelligence Scale (Resnick, 1982). Significant modifications were made. For instance, the test was lengthened and focused more on vocabulary. Terman's adaptation of Binet's IQ test quickly became popular in the United States and was not altered in any way for twenty years (White & Hall, 1980). Asa G. Hilliard III (1990) notes that Binet's original scale
was designed for "diagnostic purposes" but ended up being used for "ranking, classification, labeling, and sorting" (pp. 185-6).

The intelligence test was first developed to predict how well children would do academically in school. During WWI the role of IQ tests became even more embedded into our society. In 1917, Lewis Terman and other psychologists designed the Army Alpha Test. Over 1.7 million soldiers took the Alpha test which essentially ranked the soldiers. This group administered test was used by the military to determine who would be suitable for officer training and to cast out any men who did not meet the standards. The results of the Army Alpha showed that almost thirty percent of the men who enlisted in the army "could not read well enough to take the written form of the army tests, and of these had attended school" (Resnick, 1979, p. 246). This discovery alarmed the nation and as a result it set into motion an era of standardized testing.

After World War I, tests were the primary instrument used in assessing student knowledge and there were over two hundred achievement tests to choose from (Resnick, 1982). By 1926, the Scholastic Aptitude Test (SAT) was developed to predict which students would be successful in college (White & Hall, 1980). In 1929, the Iowa Test of Basic Skills and the Iowa Test of Educational Development was constructed. This statewide test was not mandatory but certainly helped shape the future for holding schools accountable. By the 1950's, testing became even more popular after E. L. Lindquist "... developed an electronic scoring machine that could read sheets, produce raw scores, and convert raw scores into standardized scores, all in a fraction of the time-and cost-it took teachers and scientists to conduct those tasks by hand" (Rothman, 1995, p. 38). As a result, standardized tests became the prime instrument used in assessing students and holding schools...
accountable. In a short duration, Binet’s Intelligence Test which was used to identify learning disabled students manifested into an instrument to screen and classify large groups of people.

Shortly after Binet created the IQ test, E.L. Thorndike developed scaled achievement tests to assess students in reading, mathematics, and handwriting (Resnick, 1982). Thorndike advocated using tests to evaluate student achievement (Worthen, et al., 1993). Thorndike promoted B.F. Skinner’s view of behaviorism and applied it to educational practices. Students were rewarded for correct responses and achieved learning by having information transmitted to them. Skills were broken down and taught step by step. Although Thorndike supported a stimulus response to learning he believed intelligence was inherited. Testing, therefore, could be used to scientifically determine one's individual place in society (Spring, 1997).

In the early 1900’s, around the same time that Thorndike was endorsing a behavioral approach to learning, the progressive movement emerged. John Dewey’s philosophy emphasized democracy and allowing students to take an active role in shaping their curriculum. Dewey believed the subject matter must be relevant to the children’s experiences and involve activities which will actively engage them. Jean Piaget held similar views (Dewey, 1938).

Piaget began his research on child development where Alfred Binet ended. Instead of just recording children’s responses he took it a step further by analyzing how they came to their answers. Piaget’s theory of cognitive development significantly changed educators’ views on how children learn. His theory suggests that the learning process occurs when individuals move from a state of disequilibrium. Instead of transmitting knowledge to students, Piaget believed educators should act as facilitators providing applicable learning
activities which allow children to construct their own knowledge.

Unquestionably, Jean Piaget was a key figure in educational reform, however, his research was not implemented until decades after he published it (Seymour, 1976). Even today as we approach the twenty first century, Thorndike's scientific approach often prevails over current research. Currently testing measurements are used to determine how students and schools are doing across the country. Teachers and school administrators are being held accountable and often revert back to Thorndike's behavioral approach to learning and ignore current research.

An example of this occurred in the 1970's when school officials were challenged when students scores began to drop. "... the average scores on the Scholastic Aptitude Test had declined by a total of 81 points between 1963 and 1977 - from 478 points to 429 points on the verbal portion and from 502 points to 470 points on the math portion, both on a 200 to 800 scale" (Rothman, 1995, p. 40). The results of these tests were published and used to evaluate teachers' techniques and to determine what children know. Schools across the country were being held accountable and sometimes suffered consequences if scores were not adequate. The public responded by demanding for more testing and pushing school administrators to get back to the basics.

In the social and political realm, traditional testing measures were often valued. Parents and politicians used students' scores to validate or refute teaching practices and the curriculum. Now more than ever, comparisons of students' scores were being made between school districts and states. In an effort to raise scores, teachers began using valuable teaching time to prepare students to take tests. As a result, only areas being tested were covered which limited the curriculum to the content of the test. Subjects such as science and
social studies were not getting covered adequately or were ignored completely.

Now, traditional forms of testing are used more than any other method to assess students' academic ability. Currently 127 million tests are given each year (Rothman, 1995). As testing increased in popularity concerns over their uses began to appear. Critics argued that achievement tests classified students in harmful ways and the questions were often biased against females and minorities. Educators also question whether these tests really tell us what children know and are capable of doing. As the debate over testing continues, alternative forms of assessment, such as portfolios, are being considered more closely by educators.

Entering the Twenty First-Century

According to Herman (1992) "Good assessment is assessment that is built on current theories of learning and cognition and that is grounded in futurists' and others' views of what skills and capacities students (and our society) will need for future success" (p.4). Current theorists believe, "meaningful learning is reflective, constructive, and self regulated" (Herman, p. 5). It was once believed that higher level thinking could only be obtained by first learning discrete isolated skills. Now it is believed that children best learn when they are not forced to memorize a set of facts but are instead encouraged to relate it to their own experiences by interpreting the information.

With this shift from a behavioral to a cognitive view, educators began to question the usefulness of traditional testing programs. In Accountability and Alternative Assessment: Research and Development Issues, Herman (1992) reports, "Thus the presence or absence of discrete bits of information, which is typically the focus of traditional multiple-choice tests, is not of primary importance in the assessment of meaningful learning, but rather how and
whether students organize, structure, and use that information in context to solve complex problems" (p. 5). Portfolios match more closely to current theorists beliefs since they encourage subjects to use cognitive skills and to actively construct their own knowledge. Another reason educators started considering portfolios as another option is because this assessment allows for self- and peer-evaluation. In real life situations people are usually required to work collaboratively to solve problems. Often, portfolio assessment involves students working together by finding solutions to complex problems and giving each other constructive feedback.

Portfolios

By the mid 1970's portfolios that were mainly used only by artists and photographers, began to appear in classrooms throughout the United States. Portfolios are still commonly used in both of these fields which represent the artists' finest works. These portfolios also often show growth over time and give greater insight to potential employers.

"Portfolios, coming from the word 'port', which means carry, and 'folio', which mean paper..." (Worthen, et al., 1993, p. 436) started gaining teachers' attention after Donald Graves published Writing: Teachers and Children at Work in 1983. In 1987, Nancie Atwell published In the Middle: Writing, Reading, and Learning With Adolescents. Both of these publications sparked enormous interest in using portfolios not only as an instructional tool but as a means to evaluate student achievements. Worthen, et al., best illustrate the many uses portfolios serve:

(a) an instructional linked way to assess student's performances, (b) a method of portraying a broad view of a students's achievement across a broad
range of content, (c) a way to document adequately a student’s growth in a particular content area, (d) a mechanism to enable students to identify their own strengths and weaknesses, (e) a method for encouraging students to participate in and take more responsibility for their learning, and (f) a valuable source of information for parents and administrators. (p. 441)

Instead of using portfolios within the classroom, states such as Vermont and Kentucky saw their many benefits and began using them for wide-scale assessment. In 1988, Vermont began using wide-scale portfolio assessment and Kentucky implemented them in 1990. In describing portfolios Castiglione (1996) states, “Portfolio assessments give a teacher observable, concrete illustrations of just what students are able to do for actual classroom assignments. In addition to exclusively successful outcomes, they provide illustrations of the difficulties and problems students encounter while they were learning” (p. 5). Castiglione’s description clearly highlights some of the benefits associated with using portfolios. Supporters of traditional forms of testing, however, argue that there are drawbacks to using portfolios. For instance, portfolio assessment takes time, money, and space. It is also unclear whether portfolios really assess what students know. Furthermore, many teachers are satisfied with current assessment practices, which include standardized testing, and do not wish to make any changes. Veteran teachers argue that throughout history, standardized testing has prevailed over other assessment measurements and will undoubtedly continue to be preferred in the twenty first century.
In the following chapter, mechanical aspects of portfolios will be explored as well as significant questions and concerns regarding the use of them. This paper will mainly analyze empirical and anecdotal evidence available in the K-12 educational system in hopes of determining whether portfolios are an effective alternative to traditional testing programs.
CHAPTER 3: INTEGRATIVE RESEARCH OF THE LITERATURE

Most of the research available on portfolios is generally classified in two categories. The studies usually examine portfolios as instructional tools or portfolios used for accountability purposes. There is not much empirical evidence available on the use of portfolios as instructional tools and most of the research consists of surveys and case studies. With these types of portfolios, teachers and students have much more flexibility in deciding what the portfolios consist of and how they will be scored. Portfolios used for accountability purposes have to meet certain mandated requirements set by the school district or the state. Teachers usually have very little involvement in the decision making and have to use specific rubrics for scoring the work. Most of the research done on wide-scale portfolio assessment investigates issues surrounding reliability and validity.

This next chapter will investigate both types of portfolios by analyzing research studies conducted in the last ten years. The chapter will begin by examining portfolios as instructional tools used within the classroom. Next, wide-scale portfolio assessment will be explored. Issues surrounding reliability and validity will be addressed as well as concerns over the amount of time and effort involved with implementing them.

Classroom Applications

The first two studies in this paper examines the potential benefits and advantages of using portfolios as instructional tools. Gomez, Graue, and Bloch (1991) and Moening and Bhavnagri (1996) found that implementing portfolios within the classroom can have positive effects on students' growth and attitudes toward learning.

The case study presented by Gomez, et al., (1991) also explored the
structural aspects of implementing portfolios within a classroom. The researchers investigated twelve elementary school teachers. The following section describes one teachers journey, Carolyn Benson, in implementing portfolios in a classroom.

Carolyn Benson teaches fifth grade in a large elementary school. Her writing program modeled the work of Donald Graves, Nancie Atwell, and Lucy Calkins. One hour each day was devoted to writing. Students participated in mini lessons and conferences constantly revising and editing their work.

Having students construct portfolios gives teachers as well as students a clear picture of the progress they have made. It also serves as an instructional tool giving educators insight in regards to their students' learning. This is clearly illustrated in one of the student's portfolio. Carolyn, "recognizes that keeping the portfolio allowed her to pursue questions about Jack's skills that would not have been possible if his work were not collected, preserved, and available for her - and him - to revisit (Gomez et al., 1991, pp. 623-4).

During the spring, Carolyn had a conference with three of her students: Peter, Simon, and Pao. She asked them to look through their portfolios and find any evidence showing growth in their writing. At first the boys were puzzled, but soon became very excited. All three boys noticed changes that occurred since the beginning of the school year. The most obvious changes were that their stories were longer and more in depth. Overtime they began to take risks by expanding their topics. Compiling their work allowed Peter, Simon, and Pao to acknowledge and celebrate their achievements.

Moening and Bhavnagri's (1996) "study examines the effectiveness of the showcase portfolio process as an instructional strategy to enhance the quality and quantity of students' writing products and also to motivate them to
participate in the writing process” (Moening et al., 1996, p. 180). A showcase portfolio consists of the student’s best pieces.

The portfolio intervention lasted for nine weeks and involved 18 middle to lower income first grade students from Detroit, Michigan. Before and after portfolios were implemented, students participated in a pre and post test. Four areas were measured for quality: “focus, purpose, content, and organization”. Scores ranged from “four (an excellent writing sample) to one (a poor writing sample).” A zero (an unscorable paper) was given for incomplete work. Two procedures were also used to determine the quantity of writing produced. First, "a total word count of all writing done" and second, “the average number of words per entry” were recorded (Moening et al., 1996, p. 183). Motivation was measured by how often the students used the writing center. Each day the writing center was available for approximately 45 minutes to all students. In order to be classified as using the writing center students had to be writing.

For nine weeks the students’ classroom teacher, Moening, implemented showcase portfolios. During the first week, fifteen minutes were devoted to mini lessons. These mini lessons involved students critiquing samples of writing as well as their own. Students were also required to write in their journals. Writing samples went through several stages, “students first wrote, then selected and edited pieces of writing, recopied their pieces, and finally placed their finished copy in their portfolio” (Moening et al., 1996, p. 185).

Before implementing showcase portfolios, students completed a pre assessment that lasted for three weeks. Evidence from this assessment showed that “the average number journal entries received per child was 3.72, with a range from 1 to 10 journal entries” (Moening et al., 1996, pp. 185-6). A total of 67 journals were collected.
After the showcase portfolio intervention, the same students completed a post test. The total of journal entries rose from 67 to 118. “The average number of journal entries collected per student was 6.67, with a range from 4 to 10 journal entries” (Moening et al., 1996). The quality of work “significantly” improved from a mean score of .91 to 2.02. The post assessment also showed that students visited the writing center more often. Prior to the intervention the mean score was 1.33 visits, however, after portfolios were introduced the mean score increased to 4.11.

Moening and Bhavnari (1996) concluded by saying, “The findings provide empirical data to support the opinions and claims of scholars that portfolios are not only an authentic evaluative procedure, but an equally powerful motivational approach and an effective instructional strategy” (p. 195). The researchers did point out, however, that a control group was not used and further studies need to be done. It is unclear whether the intervention of showcase portfolios increased quality, quantity, and motivation, or if it was the result of the mini lessons given during the first week, or simply the result of ordinary growth that might occur over a nine week period. Larger studies need to be done to determine if showcase portfolios can be used as instructional tools in the areas of quality, quantity, and motivation.

Newman’s (1996) longitudinal study found in addition to compiling work, portfolios can also contain students’ reflections and goals for improvement. The findings suggest that students became more active in making decisions about their education. Significant improvement in using the four management tools “resulted in more on task behavior and self-direction” (Newman 1996, p. 12). In a similar study, Brookhart (1996) provides some empirical evidence on the reliability of using annotation forms for students’ reflections.
Newman's study was conducted in hopes of determining whether portfolios help students with "planning, assessing and reflecting upon their own learning" (Newman 1996, p.14).

During the 1994-95 school year, 13 sixth grade students with "limited English language proficiency" began the study, but by the following year the class grew to 30. The school involved was located in an urban area and the students represented seven different ethnic backgrounds which included Thailand, Vietnam, Laos, China, Korea, Bosnia, and Ukraine.

The students who participated in this study were given a pre- and post-test on using the four management tools. The classroom teacher and the researcher had the students use these management tools to help assist them with the portfolio process. These four tools were, "Goal Cards, Time Management Sheets, Learning Logs, and a self evaluation checklist called the Friday Progress Report." The Friday Progress Report was a checklist that students used to record what tasks they completed. At the end of the 1994-95 school year, the Newman (1996) also compared the students' work with the 1993-94 students who did not use portfolios. The researcher referred to this group as the control group. Newman claimed the results from the longitudinal study suggest that students became more active in making decisions about their education and were more inclined to stay on task. The researcher, however, only stated the findings failing to provide specific evidence. Another problem with this study is that Newman used the students work from the 1993-94 school year as the control group. The findings would have been more stable if she would have used students from the 1994-95 school year. This would have ensured that the results gathered were based upon the use of portfolios and not other external factors. It is also unclear whether the positive effects were a result from the four
management tools, instruction, or the students' academic performance.

Through reflections and self-evaluations, both the teacher and the students could monitor their progress. Allowing students to reflect on their work and growth is one of the main advantages of using portfolios rather than traditional forms of testing.

Brookhart's (1996) study investigated whether students' reflections are "accurate and dependable representations of what the students think" (Brookhart, p. 2). Students who participated in the study filled out "Annotation Forms" for every entry that represented "curriculum objectives." The annotation forms consisted of, "... two multiple choice judgments, about the difficulty of the work sample and whether or not they would like to do more of that kind of work, and to write briefly 'why' they would or would not like to do more" (Brookhart, 1996, p. 2).

The first study was in 1994 and involved 367 first, sixth, and tenth grade students. Ten portfolios were selected using stratified random sampling. Every school from the district was represented for a total of 1,678 entries. Eighty two percent of the entries involved math while the remaining eighteen percent were based on Language Arts. The data consisted of students' annotation forms and teachers scores. The scores were based on a 4 point scale: 0 = not efficient to 4= efficient. The researcher failed to identify the scales in between which is one weakness in this study. Another problem with this study is the number of samples collected ranged from zero to thirteen. Brookhart (1996) failed to identify how many portfolios contained zero entries which might have made the sample size smaller.

Brookhart's (1996) findings claim that "The Student Annotation Form" provided a simple, useful tool for indicating student judgments about individual
work samples in a portfolio" (p. 8). Portfolios that consisted of nine entries had a reliability score of .64 for difficult work and .80 for more work. Similar results were found in the second study conducted in 1995. Portfolios that contained six to nine entries ranged from .55 to .66 while the category of more work ranged from .66 to .72. Based upon the research Brookhart (1996) concluded, "Students in-class behavior, use of lesson and study time, is consistent with their annotations" (p. 8). Brookhart’s research does provide some empirical evidence on the reliability of students’ annotations to their work but because of the weaknesses found in the study, the results are questionable.

One concern educators have with portfolios is the authenticity of students’ work. Teacher participation in constructing portfolios varies considerably and, therefore, the students’ entries may not be a true representation of what they know. This issue of validity is raised in Gearhart, Herman, Baker, and Whittaker’s study (1993) on portfolio assessment.

In 1991, nine teachers participated in the study. They taught at the same school ranging from grades one to six. Each teacher selected six students. Students were divided evenly into three categories based upon their writing ability: “high, medium, and low.” The teachers were instructed to help each student participating in the study to complete a portfolio as well as documenting any guidance they gave. The guidance received was rated from zero to three in the following categories: “Content-organization, Style, and Mechanics. Challenge, Copied Work, and Time was also recorded.”

The researchers analyzed the ratings of 228 assignments. The results showed students classified as having “medium” to “low” writing abilities received more instructional help than “high” students. As stated earlier, teachers rated their assistance on a scale of zero no support, to three, detailed help. In
the area of Content-organization the mean score was 1.47 for "high" students, 1.88 for "medium" students, and 2.05 for "low" students. The category of Mechanics and Style also had similar results suggesting that "high" students do not receive as much help compared to "medium" and "low" students.

Another interesting finding was that teachers who had more experience using portfolios tended to assist their students more. Teachers with more experience using portfolios ranged from a mean score of 1.57 to 2.12. While teachers with less experience averaged from 1.37 to 1.63. This evidence indicates that the amount of assistance teachers give varies. The researchers concluded, "Thus the validity of inferences we can draw about student competence based solely on portfolio work becomes suspect" (Gearhart et al., 1993, p.7).

In a survey conducted by Egelson (1995) teachers complained that the younger students needed much more help in constructing their portfolios. At the end of the 1993-94 school year 11 teachers in a Southeastern elementary school answered four open-ended questions. The results from the survey showed that the level of teacher involvement fluctuated greatly based upon the grade level of the child. Younger children received more assistance and guidance than the older children.

If portfolios are going to be used for accountability purposes then teachers' participation and guidance needs to be consistent between districts and states. Currently the amount of assistance teachers give in helping students construct their portfolios varies considerably. More studies need to be done on the effects of different approaches teachers take and the amount of assistance students receive outside of the classroom. Otherwise, the contents found in the portfolios may not accurately represent student achievement.
Although the studies are small and usually lack empirical evidence, the research implies that there are benefits to using portfolios. As an instructional tool, portfolios can provide teachers with detailed information about students' strengths and weaknesses. However, it is important to take into account that many of the studies had flaws which may have skewed the results. To validate the findings more studies with larger sample sizes need to done and more empirical evidence needs to be provided especially when used for accountability purposes.

The first section in this chapter explored the use of portfolios as instructional tools within the classroom. The following section will explore significant questions and concerns regarding the use of wide-scale assessment. First, however, this paper will discuss the mechanical aspects of implementing portfolios in large-scale applications.

Wide-Scale Applications

Through reflections and self-evaluations, portfolios actively involve students in their education. Three teachers interviewed by Irwin-DeVitis (1996) were concerned, however, that the benefits associated with using portfolios within the classroom would be lost if wide-scale portfolio assessment were to be implemented. Although this is not an empirical study, it provides insight into the applications of individual classroom and wide-scale portfolio assessments.

All three teachers used literacy portfolios in their classrooms. They taught at different schools but within the same district. The three teachers, Lynda, Ed, and Rolanda regularly met with a study group to share experiences and seek support from one another. They define literacy portfolios as "a collection of work that demonstrates the student's ability to read and write in a wide range of contexts and for a variety of purposes, to document the processes used, and to
evaluate his or her own efforts" (Irwin-DeVitis, 1996, p. 225). The three teachers value literacy portfolios because they give students the opportunity to reflect on their own growth and take risks. Students select entries that are meaningful to them and illustrate their literacy skills.

For example, a fourth grade student, Denise, chose to include a telephone bill along with literature on phone rates in her portfolio. Being able to decipher the phone bill and rates became very important to Denise after her best friend moved to another state. Denise’s parents agreed to pay $15.00 worth of the long distance phone calls to her friend. Additional costs per month had to come out of Denise’s allowance. As a result, Denise became very familiar with different phone rates and what time of the day is best to call. This real life experience, documented in her portfolio, demonstrates her ability to use her literacy skills effectively. Giving students ownership encourages creativity and diversity which is lacking in most testing programs.

These three teachers are committed to using literacy portfolios because they capture what students can actually do. Lynda sums it up best by saying, “Portfolios are time-consuming, but they give the most accurate picture of the student. I can see growth over time in real and relevant reading and writing activities” (Irwin-DeVitis, 1996, p. 229).

As their district was moving toward large scale portfolio assessment, these three teachers had doubts of its success. Unlike the literacy portfolios used in their classrooms, these mandated and standardized portfolios would have to meet certain requirements. It would be highly unlikely that with these restrictions that Denise would be able to include a phone bill and literature on phone rates in her portfolio. As a result, students would lose ownership over their portfolios and creativity and diversity would be stifled. The author concludes that these
three teachers, "have grave doubts that the district standardized portfolio assessment plan will be able to maintain the authenticity that they consider so critical in portfolio assessment" (Irwin-DeVits, 1996, p. 233).

The goal of The National Standards Project is to establish a "national examination system" which includes the use of portfolios. The founders of the project Lauren Resnick (1991) and Mark Tucker (1991) hoped to create a system that was "syllabus driven" and addressed concerns like the ones proposed by the three teachers Lynda, Ed, and Rolanda. Teachers would act more as "coaches" rather than as "evaluators" by helping students prepare for examinations. The examinations would consist of "projects, portfolios, and performances." States or several districts would determine the tasks to be completed. The national examination system would only provide assistance and support. It would not dictate the content of the portfolios. However, states or school districts could set mandated standards which is what Lynda, Ed, and Rolanda from Irwin-DeVitis's (1996) study were afraid of.

If the National Standards Project was put into effect, states and districts within states would have different requirements and scoring systems. Just as with traditional testing programs, the scores would be used for college admissions and to evaluate how well schools are doing. Unlike traditional testing measurements, the National Standards Project would not establish a uniform system. States or school districts would decide what the requirements for the portfolios would consist of. The question then remains, without a uniform system would cross-state comparability be possible?

Linn, Kiplinger, Chapman, and LeMahieu (1992) explored this question. During the summer of 1991, representatives from ten states Arizona, California, Colorado, Connecticut, Maine, New York, Oregon, South Carolina, Texas, and
Vermont attended a "cross-state scoring workshop" for four days. Each state was asked to send three representatives: "a state-level writing curriculum supervisor, a state-level evaluation expert, and current classroom teachers experienced in the state writing examinations." Not every state was able to do this. Each of the ten states were instructed to send writing samples with attached scores. Representatives also brought literature explaining their program and scoring systems. At the workshop representatives from each state scored writing samples from other states. As a result, each writing sample had a total of 4 ratings. Its original rating plus three more from different states. Forty eight correlations were obtained from 8 states submitting elementary students' writing. Representatives were instructed to use the scoring system they were familiar with and were not allowed to see previous scores.

For the purpose of this study a total score was used for cross-state comparisons. The results show a high level of agreement between raters from different states. At the elementary level cross-state correlations reached as high as .87. The mean correlation was .73. The researchers conclude, "The high correlations among the scores assigned by readers from different states were achieved without modifications of the scoring procedures used by the 10 states participating in the workshop" (Linn et al., 1992, p. 107). Rating agreement between the states was high despite the fact a variety of tasks at different grade levels was examined. According to this study, cross-state comparability would be possible and using different scoring procedures does not change the outcome of the score. However the issues Lynda, Ed, and Rolanda raised still need to be addressed. More studies with larger size samples need to be conducted to determine the mechanical aspects of wide-scale portfolio assessment, especially in regards to mandated standards.
Reliability and Validity

Another concern with wide-scale portfolio assessment has to do with reliability. This next section investigates empirical evidence on interrater reliability. If portfolios are going to be used to assess students then there must be agreement between raters on scoring individual items or the portfolio as a whole. Herman and Winter's (1994) propose some very important questions, “Do raters agree on how a portfolio ought to be scored? Do they assign the same or nearly similar scores to a particular student's work” (p. 49)? The following studies explore these questions.

In 1988, Vermont began developing a statewide assessment program that was primarily portfolio based. Rick Mills, the Commissioner of Education, and others helped create an assessment program they hoped would “provide useful information about student performance, and to encourage improvement in teaching” (Koretz, et al., 1992, p. 12). Researchers from the Rand Institute on Education and Training National Center for Research on Evaluation Standards, and Student Testing (RAND-CRESST) investigated the Vermont Portfolio Program. The following section focuses on Koretz et al.,’s analysis and recommendations concerning reliability and validity.

This wide-scale assessment program was designed to be used for many subjects and grades but was only implemented in grades four and eight in writing and mathematics. Many educators, however, broadened the program by including other grades and subjects. Results from these areas, however, were not included in the study.

The mathematics portfolio consisted of five to seven pieces that represented “puzzles, applications, and investigations.” Each example was scored on seven dimensions and each dimension was scored on a four point scale. Four pieces
represented problem solving while the remainder were examples of communication. The students' portfolios were not rated by their own classroom teachers but by volunteer teachers. In most cases the contents in the portfolios were graded twice by different raters.

During the 1991-92 school year, the five researchers investigated the results of 1,247 math portfolios. The results indicated that the scores between the two raters were not in agreement. Reliability coefficients ranged from .23 to .45. Based upon the findings the rater reliability was extremely low, and, therefore unreliable.

The results from the writing portfolio showed similar results. Unlike the math portfolio, the writing portfolio consisted of two parts. Students were required to include six to eight pieces that represented “a poem, short story, or narration and a personal response to a book, event, current issue, mathematical problem, or scientific phenomenon” (Koretz et al., 1992, p. 13). Students selected one piece from each category that was considered to be their “best piece.” The remaining examples were then scored as a set rather than individually. Each “best piece” and “writing remainders” were scored on five dimensions which included: “purpose; organization; details; voice-tone; and Usage-mechanics-grammar.” Each dimension was scored on a four point scale: “1. rarely; 2. sometimes; 3. frequently; 4. extensively.”

Another difference in the writing portfolios was the classroom teachers graded their own students’ examples while volunteer teachers scored the contents in the math portfolios. Although the researchers claimed the results from the study suggest teachers do not favor their own students' portfolios when scoring them they should have followed the same procedures for the mathematics portfolios. This would have eliminated any suspicion on the
findings.

Like the results from the math portfolios, reliability was generally low. In the writing remainder section, which included all the pieces except for the best piece, the reliability coefficients ranged from .28 to .57. The best piece section also produced similar results with an average of .35 in grade four and .42 in grade eight.

The data in this study found that the raters scores from both mathematics and writing were not reliable. The portfolios could not be used to accurately measure the quality of work produced by the students since the agreement between the raters was low. RANDS-CRESST's researchers concluded that "inadequate scoring rubrics, insufficient training, and the lack of standardization of tasks contributed to the low reliability of the scores" (Koretz et al., 1992, p. 18). Since the reliability of the scores were so low the issue of validity was not achieved.

During the 1992-93 school year many adjustments were made to help improve rater agreement. In June, all of the portfolios were scored at one location over a period of five days. Raters also attended workshops which involved grading pieces and then discussing the results. Even with these changes, reliability still remained a problem especially with the writing portfolios. In the category of best pieces the correlation score rose from .35 in 1992 to .40 in 1993. In grade eight the correlation increased from .42 to .45. The data clearly illustrates that not much progress was made during the second year.

Rater reliability in the mathematics portfolios, however, fared much better. Koretz, Klein, McCaffrey, and Stecher (1993) concluded that the mathematics portfolios looked "encouraging" but there was still "room for improvement"
(Koretz et al., 1993, p. 11). With these encouraging results also came recommendations to further explore issues regarding validity.

In regards to the writing portfolios, however, the RAND-CRESST’s researchers concluded that drastic changes needed to be done in order to increase interrater reliability such as: “more conventional (and narrower) definitions of type of tasks; and rubrics that are simpler and perhaps genre-specific” (Koretz et al., 1993, p. 12).

Novak, Herman, and Gearhart (1996) conducted a comparative study to determine the reliability and effectiveness of a rubric called Writing What You Read (WWYR). Unlike other rubrics used for scoring the WWYR consisted of “five analytic subscales for theme, character, setting, plot, and communication, and a sixth holistic scale for Narrative Effectiveness” (Novak et al., p. 3).

This study compared results from the sixth scale of the WWYR and another rubric currently being used for wide scale narrative assessment. Five raters were familiar with both the WWYR and the comparison rubric. The results from the study showed that the WWYR rubric produced more reliable scores. Between the five raters the mean correlation for the WWYR rubric was .69 while the comparison rubric was .45. The research indicates that wide-scale narrative assessment is possible if a rubric such as the WWYR is used. The authors conclude by saying “... our findings suggest that choice of rubric can have substantial effect on both technical quality and the results of a performance assessment” (Novak et al., 1996, p. 28). The authors warn, however, that “strong inferences are not warranted” since the size of the sample was small but the results do “hold promise for the pursuit of large scale writing assessments” (Novak et al., p. 28).

LeMahieu, Gitomer, and Eresh (1995) do not believe that a rubric alone
improves interrater reliability. These three researchers believe the success of Pittsburgh's Portfolio Program is a result of "teachers and administrators repeatedly examining student work and developing a vocabulary that people could use to discuss that work" (p. 25). According to this study, wide-scale portfolio programs can produce reliable scores if there is a "shared understanding within the educational community" (p.25).

In 1992, writing portfolios were implemented in the Pittsburgh Public School District. Stratified sampling was used in selecting 1,250 portfolios to be scored during a designated week during the summer by 25 raters. The portfolios ranged from grades six through twelve and were scored on three dimensions: "1) Accomplishment in Writing, 2) Use of Processes and Strategies, 3) Development as a Writer." Each dimension was rated on a six point scale which ranged from "Inadequate Performance" to "Outstanding Performance." Students selected four items that they felt represented their writing ability. These four items were categorized as: "1) an important piece, 2) a satisfied piece, 3) an unsatisfied piece, and 4) a free pick'. A free pick is an additional piece that the student chooses. Teachers could also include a fifth entry if they believed the portfolio did not fully represent their students' writing abilities. The raters could also mark a dimension as "No Evidence" (NE) if they felt it lacked enough evidence to make a judgment. In addition, the portfolios also contained drafts and reflections on the chosen selections.

Each portfolio was scored twice out of a group of twenty five skilled raters. Portfolios were scored by a "Chief arbiter" if "the judges ratings differed by more than one point, or, if one judge had an NE rating an one did not" (LeMahieu et al., 1995, p. 14). The researchers found that the agreement among raters was high. At the middle school level, the reliability scores ranged from .87 to .84.
while at the high school level it ranged from .80 to .75.

Compared to the Vermont study, Pittsburgh's reliability scores were significantly higher. The data obtained from the Pittsburgh study is better since stratified sampling was used. Portfolios from all over the district were scored which represented 8% of the student population. The researches also did not have the students' teachers score their work which occurred in the Vermont study.

The Vermont researchers believe agreement between the raters could have been improved with a better developed rubric. Although the study conducted by Novak, et al., (1996) is small their findings suggest that the WWYR rubric could be used effectively in a wide-scale assessment. The Pittsburgh researchers, however, warn, "It is, in our view, a misguided notion to seek solutions to the reliability problems exclusively in the particulars of any rubric" (LeMahieu et al., 1995, p. 25). Instead, more effort should be focused on creating a common understanding between educators and administrators.

A case study conducted by Daniels (1995) illustrates the tension one Vermont teacher felt when asked to implement portfolios in her classroom. In 1995, Vermont switched from assessing fourth grade students to assessing fifth grade students using portfolios. This change affected Maura Lincoln, an experienced teacher of twenty years. During a ten month period, Daniels observed and interviewed Lincoln. She had a total of 23 fifth grade students, seven of whom required Individual Education Plans (IEP).

Daniels noted that Lincoln felt at times frustrated and wanted desperately to be given detailed instructions on what to do. She did not use portfolios effectively as an instructional tool because she was so concerned about the end product. Lincoln explains, "It would have been nice if I could have a year to
prepare and get ready. Then I'd be ready for it. Then I wouldn't have this pressure of 'What do they want? What will it look like in the end?' (Daniels, 1995, p. 14).

Daniels study gives some evidence that communication between teachers and administrators was lacking in the Vermont program. Lincoln was at a loss on how to implement portfolios in her classroom and did not feel she was given much support and guidance from administrators. More studies need to be conducted to determine if the agreement between raters increases based upon the rubric being used as proposed by the Vermont researchers, or if it has to do with teachers and administrators collaboratively working together as suggested by the Pittsburgh researchers.

Another important characteristic associated with educational assessment has to do with validity. Herman and Winter's (1994) explain, "When we assess students what we really want to know is, do the scores of a portfolio assessment represent some enduring and meaningful capability? Are scores good indicators of what we think we are assessing" (p. 51)? One way to determine this is by comparing results from two measurements. In two studies, one conducted by Stevens and Clauser (1996) and the other by Benoit and Yang (1996) the results indicated that the portfolio assessment did not correlate with the traditional standardized tests used. The following section explores whether portfolios overestimate, underestimate, or give accurate estimates of students' skills by comparing the results with traditional standardized tests.

The following four-year study was administered to compare results from the Iowa Tests of Basic Skills (ITBS) with the New Mexico Writing Portfolio Assessment (WPA). From the spring of 1992 to the spring of 1995, 2,351 students from a “large Southwestern school” participated in the study. Students
were required to take the ITBS twice. Once in the spring of 1992, and again in 1994. Vocabulary, reading, spelling, capitalization, punctuation, and language use and expression was assessed though mainly multiple choice type questions.

The same students also completed New Mexico’s Writing Portfolio Assessment (WPA) during the spring of 1993 and 1995. Unlike other portfolio assessments, the WPA required students to respond to three prompts. Out of the three prompts one was sent to be scored by a contractor. The state chose which prompt would be examined. The pieces were scored on four scales: “Development, Word Usage, Sentence Formation, and Language Mechanics.”

Stevens and Clauser (1996) only examined the scores of students who participated all four years by completing both assessments. The ITBS scores showed higher intercorrelation with one another than the WPA. In the ITBS, third grade correlations were .694 and the average fifth grade correlations were .714. The WPA “average correlations of the scores in Grade 4 was .507 and in Grade 6 was .570” (Steven & Clauser, p. 6). WPA predictive validities were also lower than the ITBS. The predictive validities in the WPA ranged from .301 to .188 while the ITBS predictive validities ranged from .763 to .559.

In a similar study by Benoit and Yang (1996) results from the Texas Assessment of Academic Skills (TAAS) was compared with scores given on portfolios. In the spring of 1992, the Texas Education Agency required Chapter 1 instructors from the Dallas Independent School District to include the use of portfolios to assess their students. The portfolios included a “district-developed checklist” and at least six samples of students’ work.

The selections were scored on a 3 point scale ranging from “not yet =0, some of the time =7, and most of the time =10. Individual scores were then combined
to create a total score. Unlike other studies examined so far, only the classroom teachers scored their students' portfolios. The issue of interrater reliability, therefore, could not be analyzed.

In conjunction with portfolios, students were also required to take the TAAS which consists of multiple choice questions. Results obtained from third grade Chapter 1 portfolio ratings were compared with scores obtained on the TAAS during the 1993-94 school year.

The researchers found that "Overall, the results indicate a positive, but weak, association between the two measures" (Benoit & Yang, 1996, p. 185). Portfolio ratings were higher at the "most of the time" level totaling 44% while the TAAS scores totaled at 33%. However, at the "not yet" level portfolio ratings were much lower than the TAAS scores (2% versus 11%). Overall, the scores between the two assessment measurements did not correlate. Based upon these findings it is in the opinion of the authors that mandated standards be set in order to insure the information obtained is valid and reliable and portfolios used for "accountability purposes" must have a "clearly defined criteria and appropriate rubrics" (Benoit & Yang, 1996, p. 190). Advocates for portfolios, however, argue that scores obtained from standardized tests are not accurate either. More studies on both portfolios and traditional testing programs need to be conducted to determine which measurement provides the most accurate information about students' abilities.

In their conclusion, Benoit and Yang (1996) make an interesting point, "It is time to recognize that there are truly two types of portfolio assessments: an accountability model and a classroom model" (p. 190). Castiglione (1996) holds this same view. After comparing standardized tests with portfolios, he concludes, "one cannot, very well substitute one for the other without changing
measurement targets and outcomes” (p. 10). In his opinion they both serve different purposes. For large scale assessment, standardized tests are better equipped than portfolios. However, portfolios are more beneficial for classroom settings. In other words, “you cannot utilize a single portfolio assessment system to accomplish both goals” (Benoit & Yang, p. 190).

Would it be feasible to implement two portfolio systems one for “accountability purposes” and one to “improve instruction and learning?” According to Roe and Vukelich (1994) in “Portfolio Implementation: What About R for Realistic?” the answer is no. This pilot project investigated challenges and benefits primary grade teachers faced when implementing portfolios in their classrooms. Ten teachers were invited to participate in the pilot conducted by the University of Delaware. The ten teachers were chosen by their principals. Out of the ten, seven teachers participated. The authors do not specify how many schools were involved but claim they covered a diverse population.

The teachers were required to meet any state and district mandates as well as using portfolios. The findings obtained from the surveys showed that having to complete both forms of assessment were overwhelming for the teachers. One teacher complained, “Teachers cannot continue to do district required assessments and the portfolio assessment piece. It put a great deal of pressure on us” (Roe & Vukelich, 1994, p. 13). The pressure would have been even greater if the teachers had to implement two portfolio systems as suggested by Benoit and Yang (1996), and Castiglione (1996).

Roe and Vukelich also found that when teachers had to keep portfolios as well as meet any state and district requirements the easier system prevailed. The amount of time involved using portfolios is far greater than most testing programs available. As a result, Roe and Vukelich (1994) found that teachers
did not use portfolios to their potential. Instead they became more like work folders which contained work but was not "a personal documentation of each student's literacy development" (p. 12). As stated earlier, a disadvantage of using portfolios compared to existing testing programs is that they involve much more time and effort. However, when used correctly, portfolios can provide detailed information about students as learners which is more difficult to achieve with standardized tests. The question remains whether the benefits of using portfolios are worth the extra time, effort, and money involved with implementing them. According to the The General Accounting Office, alternative testing methods, such as portfolios, cost $33 per student while multiple-choice tests cost less at $16 per students (Rothman, 1995). The following studies explore teachers' perspectives on the advantages and disadvantages of using portfolios.

Advantages and Disadvantages

Shapley and Pinto (1997) examined the reactions of the teachers in the Independent Dallas School District had toward using a large-scale portfolio assessment. Earlier, Benoit and Yang's (1996) study on the validity of Dallas's portfolio program was discussed. In 1992, the Dallas Independent School District Chapter 1 teachers were required to include using portfolios to assess their students in grades K-3. In the 1994-95 school year all of the Chapter 1 instructors were invited to participate in the study.

Two surveys were given during the 1994-95 school year. The first survey was an interview which took place in the fall. Trained monitors asked all of the participants five open-ended questions. The second survey occurred in the spring. It was a questionnaire that was completed by 93 out of the 101 Chapter 1 teachers. The questionnaire consisted of 24 questions which could be
answered from a range of one to five. One being “not true of me now” to five being “very true of me now.”

The results from the data collected from the questionnaire concluded that during the 1994-95 school year the teachers had a positive experience using portfolios and found them beneficial in assessing their students' strengths and weaknesses. The data, however, also showed that when the portfolios were first implemented in 1992 the teachers encountered many difficulties, but felt they were able to overcome them through adequate training. Out of 116 surveyed 23 responded “not true of me now” to the category of “Portfolio assessment requires more of my time than it is worth.” Only 10 responded to the same category “very true of me now.” Although more teachers value portfolios, there are still some teachers who clearly do not.

In another survey Frederick (1996) found that although portfolios are gaining popularity, teachers still value traditional forms of assessing students' skills. The participants were asked to “include demographic information, the amount of training in portfolio use, and what perceived strengths, weaknesses, benefits, and barriers they feel are associated with portfolio use” (Frederick, p. 2). The entire questionnaire took approximately ten minutes to fill out and the participants remained anonymous.

Out of the 162 surveyed, 80% of the teachers used portfolios in their classrooms, however, 90% “felt that report cards would not be replaced by portfolios” (Frederick, 1996, p. 5). The areas that teachers used portfolios for assessing students the most was in Language Arts (36%) and in Reading (16%). Only 16% used portfolios in every subject. Strengths associated with portfolios included showing students progress (47%) and encouraging students to reflect on their work (15%). Weaknesses the teachers addressed were
portfolios are too time consuming (46%) and lack of training (23%). The evidence gathered from this survey suggests that although portfolios are gaining popularity, teachers still value traditional forms of assessing students' skills.

Calfee and Perfumo (1993) found similar results in their survey. The participants were not randomly chosen. The survey consisted of four starter questions: “1) How did you get into portfolios? 2) What does the concept mean in practice? 3) How do you do it? and 4) What do you see as the effect of portfolios for your students” (p. 533)? To obtain further information, 24 respondents participated in a 2-day conference. The article does not specify how the respondents were selected. Group sessions were video-taped. Graphic reports and individual reflections were also complied.

As a result, 70 packets were collected which represented two state-level projects and several districts perspectives on the use of portfolios. The results indicated that overall educators were pleased with the process and commented how students gained a sense of ownership and pride with their work. Teachers indicated they were uncomfortable with assessing the portfolio as a whole but did not mind critiquing individual items. Rather than assigning grades many teachers chose to point out students' strengths and weaknesses instead.

The findings based upon all three studies show that portfolios have many difficult challenges. They take up a considerable amount of time and effort compared to most testing programs. Nevertheless, even with these challenges many teachers were pleased with the portfolio process. However, when examining the findings presented here, it must be taken into consideration that in all three surveys the sample size was small and stratified sampling was not used. To support these findings more empirical evidence needs to be
conducted using larger size samples.

It is clear that in both realms, within the classroom and in wide-scale applications, more studies on portfolio assessment needs to be conducted. Issues surrounding "technical foundations" and "evaluation standards" must be addressed if portfolios are going to make a permanent mark in our educational system. Teachers need additional support as well as technical assistance in implementing portfolios both as instructional tools and for accountability purposes. If these issues are addressed portfolio assessment may have a promising future in our educational system. Otherwise, traditional testing programs will prevail.
CHAPTER 4: CONCLUSION

This paper analyzed research available on portfolios used for educational purposes. Empirical and anecdotal evidence was provided in hopes of answering the question of: What are the mechanical aspects of portfolios and what has research shown in terms of significant questions and concerns regarding the use of them?

The third chapter began by exploring the benefits and advantages of using portfolios as instructional tools within the classroom. The findings suggest that classroom portfolio assessment is more closely linked to classroom instruction. The portfolio process also gives students the opportunity to reflect and celebrate on their achievements. Empirical evidence was provided on the reliability and validity of students’ reflections using management tools and annotation forms as well as the use of portfolios as instructional tools. However, Herman, Baker and Whittaker’s (1992) study Whose work Is it? raises concerns over the authenticity of students’ portfolios and whether they really assess what students can do.

Next, research on wide-scale portfolio assessment was examined. Although difficult and time consuming, Linn, et al.’s (1992) research suggests that cross-state comparability is possible. It is also unclear whether wide-scale portfolio assessment can maintain the special qualities found within the classroom once mandated standards are set. More studies need to be conducted that address these concerns.

Issues surrounding reliability was also investigated. The results on interrater reliability was mixed. In the Vermont experience, the reliability of the scores were low and, therefore, could not be published. The Pittsburgh experience had much better results in achieving high interrater agreements. Two other studies
were analyzed that examined whether portfolios produce similar results to standardized tests. In both studies, the results from the portfolio assessment did not correlate with the traditional standardized tests used.

Lastly, this paper explored the advantages and disadvantages involved with implementing portfolios. The studies suggest that although portfolio assessment entails many challenges, the potential benefits and advantages outweigh the additional time and effort involved with implementing them. Many of the studies presented in this paper, however, had flaws or weaknesses, and, therefore, the findings cannot be supported.

The question of whether one portfolio assessment can be used for both accountability purposes and as an instructional tool is still not answered. Asking teachers to implement two separate portfolio assessments appears unrealistic, however, at the same time, trying to combine two assessments into one seems equally as impossible.

In many of the studies examined in this paper, concerns over reliability, validity, and mandated standards have been raised. Additional research needs to address these concerns by providing more empirical data and larger studies that include stratified sampling and control groups. Then, perhaps, the question of whether one portfolio system can meet the needs of both the "accountability model" and the "classroom model" can be answered. If further research shows that one portfolio instrument cannot serve both purposes then school officials, educators, and parents must decide whether or not to implement both standardized tests and portfolios or to choose between the two.

As it stands the future of portfolio assessment is not clear. Calfee and Perfumo (1993) suggest that the portfolio movement could follow three different paths: "1) it may disappear for lack of audience, 2) it may become standardized,
or 3) it may become a genuine revolution" (p. 536). Due to a lack of empirical research and adequate technical training, portfolio assessment may never fully materialize in our educational system. Wortham (1995) quotes her colleague who states educators are, "takin' the talk, but not walkin' the walk." Wortham compares the popularity of portfolios to the concept of open classrooms which started its momentum in the 1970's. She argues that the concept of open classrooms were immediately put to use and then quickly discarded because educators were inexperienced and had little guidance. Wortham warns that portfolios will have the same fate as open classrooms if teachers are not properly trained in using portfolios for assessment. Without adequate training and support, educators will most likely get frustrated and resort back to traditional forms of testing and grading.

If more and more teachers receive adequate training and support, it is possible, especially with the popularity portfolios are receiving, that a national wide-scale portfolio system be put into place. However, with concerns over reliability and validity this system may become too standardized. In Linda Irwin-DeVitis' (1996) study, the three teachers interviewed were concerned that the mandated and standardized portfolio system their school district proposed did not resemble the literacy portfolios they were currently using in their classrooms. Portfolio programs that have strict guidelines and requirements may diminish the positive characteristics found in them. Students and teachers would not have as much freedom in constructing the portfolios which might in turn destroy the authenticity associated with using them. Issues regarding validity and reliability must by addressed without sacrificing the special qualities found in using portfolios for assessment. Portfolio assessment in individual classrooms, however, may flourish. Constructivist teachers in particular may rely
entirely on portfolio assessment only using traditional testing measurements when required to.

The last possibility proposed by Calfee and Perfumo (1993) is that the portfolio movement will “become a genuine revolution” (p. 536). Portfolios would completely reform and restructure our educational system by replacing current testing programs that include tests such as the norm-referenced test.

Historically standardized tests have prevailed over other forms of assessments. During the Progressive era’s, traditional testing measurements were challenged by more authentic assessments such as portfolios. In the end, however, multiple choice type tests were more widely accepted in both the political and social realms. Recently, the same ideals found in the Progressive era’s, are resurfacing putting portfolio assessment in the spotlight. Before portfolio assessment is accepted, however, more evidence needs to be provided that proves it is indeed the best assessment program available. Until then, standardized tests will most likely remain the most used assessment method in our K-12 educational system.
References


Nygren, B. (1978). Was This What You Really Had in Mind, Horace Mann?
Journal of Teaching and Learning. 3, 46-60.


Association, Chicago.


