

Lindenwood University

Digital Commons@Lindenwood University

---

Dissertations

Theses & Dissertations

---

Spring 5-2009

## Middle School Grading Practices and the Ability to Predict Achievement on the Arkansas Benchmark Test

Philip Matthew Summers  
*Lindenwood University*

Follow this and additional works at: <https://digitalcommons.lindenwood.edu/dissertations>



Part of the [Educational Assessment, Evaluation, and Research Commons](#)

---

### Recommended Citation

Summers, Philip Matthew, "Middle School Grading Practices and the Ability to Predict Achievement on the Arkansas Benchmark Test" (2009). *Dissertations*. 618.

<https://digitalcommons.lindenwood.edu/dissertations/618>

This Dissertation is brought to you for free and open access by the Theses & Dissertations at Digital Commons@Lindenwood University. It has been accepted for inclusion in Dissertations by an authorized administrator of Digital Commons@Lindenwood University. For more information, please contact [phuffman@lindenwood.edu](mailto:phuffman@lindenwood.edu).

Running head: MIDDLE SCHOOL GRADING PRACTICES

Middle School Grading Practices and the Ability to  
Predict Achievement on the Arkansas Benchmark Test

by

Philip Matthew Summers


May, 2009

A dissertation submitted to the Education Faculty of  
Lindenwood University in partial fulfillment of the  
requirements for the degree of  
Doctor of Education  
School of Education

DECLARATION OF ORIGINALITY

I do hereby declare and attest to the fact that this is an original study based solely upon my own scholarly work at Lindenwood University and that I have not submitted it for any other college or degree here or elsewhere.

Full Legal Name: Philip Matthew Summers

Signature:  Date: 8-16-09

MIDDLE SCHOOL GRADING PRACTICES AND THE ABILITY TO  
PREDICT ACIEVEMENT ON THE BENCHMARK TEST

by

Philip Matthew Summers

A Dissertation has been approved as partial fulfillment of  
the requirements for the degree of  
Doctor of Education  
at Lindenwood University by the School of Education

  
\_\_\_\_\_

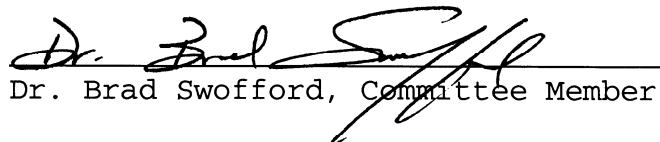
Dr. Terry Reid, Dissertation Chair

Aug 10, 2009  
Date

  
\_\_\_\_\_

Dr. Sherry DeVore, Committee Member

Aug 10, 2009  
Date

  
\_\_\_\_\_

Dr. Brad Swofford, Committee Member

8-10-09  
Date

## ACKNOWLEDGEMENTS

Thank you to the following people for encouragement, support, and valuable insight throughout the entire dissertation process: Dr. Terry Reid, Dr. Sherry DeVore, Dr. Brad Swofford, Mrs. Lucy Lyon, and Ms. Joana King.

## Abstract

After the No Child Left Behind Act was legislated, it became necessary for states to target specific learning goals and then test those objectives. As part of this process, districts began to develop new curricula and evaluate grading practices. For this study, student samples were drawn from sixth, seventh, and eighth grade populations of a Northwest Arkansas Middle School. Samples were disaggregated by grade level for the 2005-2006, 2006-2007, and 2007-2008 school years. A quasi-experimental design was implemented to test the strength of the independent variable, averaged semester grades, on the dependent variable, Arkansas Benchmark Test scores. A Pearson  $r$  correlation was the primary measurement tool and the coefficient was calculated for each grade level for each of the three years. The results showed no statistically significant link between the averaged semester grades of the Northwest Arkansas Middle School and the Arkansas Benchmark scores.

TABLE OF CONTENTS

LIST OF TABLES.....vii

KEY TO ABBREVIATIONS.....viii

CHAPTER ONE - INTRODUCTION.....1

    Background.....1

    Theoretical Underpinnings.....5

    Statement of the Problem.....6

        Purpose of the Study.....8

        Questions.....9

        Independent Variable.....9

            Semester Literacy and Math Grades.....9

        Dependent Variable.....10

            Arkansas Benchmark Test Scores.....10

    Hypotheses.....10

        Null Hypothesis.....10

        Alternate Hypothesis.....10

    Limitations of the Study.....10

    Definitions of Terms.....11

    Summary.....14

CHAPTER TWO - LITERATURE REVIEW.....16

    Background.....16

        History.....16

    Purpose of Grading.....24

    Theory.....27

Reform.....	39
Effects.....	61
Grading Practices.....	61
Mandated Testing.....	67
Summary.....	78
CHAPTER THREE – DESIGN AND METHODOLOGY.....	80
Introduction.....	80
Subjects.....	80
Sampling Procedure.....	82
Research Setting.....	83
Research Design.....	83
Questions.....	84
Independent Variable.....	84
Semester Literacy and Math Grades.....	84
Dependent Variable.....	85
Arkansas Benchmark Test Scores.....	85
Hypotheses.....	85
Null Hypothesis.....	85
Alternate Hypothesis.....	85
Procedure.....	85
Statistical Treatment of Data.....	87
Pearson r Correlation.....	87
Coefficient of Determination.....	88
Summary.....	88



CHAPTER FOUR - RESULTS.....	90
Introduction.....	90
Results.....	90
Analysis of Data.....	90
Research Question Number One.....	90
Research Question Number Two.....	92
Pearson $r$ .....	92
Coefficient of Determination.....	94
Research Question Number Three.....	96
Survey.....	96
Deductive Conclusions.....	100
Summary.....	101
CHAPTER FIVE - DISCUSSION.....	102
Introduction.....	102
Implications for Effective Schools.....	104
Recommendations.....	104
Summary.....	105
REFERENCES.....	106
APPENDIX A.....	120
APPENDIX B.....	130
<i>E-mail to Area Educators</i> .....	130
VITA.....	131

LIST OF TABLES

Table 1. *Demographics*.....81

Table 2. *Correlations for 1<sup>st</sup> and 2<sup>nd</sup> Semester Literacy Grades*.....91

Table 3. *Correlations for 1<sup>st</sup> and 2<sup>nd</sup> Semester Math Grades*.....92

Table 4. *Correlation of Averaged Semester Grades and Benchmark Raw Score Percents Literacy*.....93

Table 5. *Correlation of Averaged Semester Grades and Benchmark Raw Score Percents Math*.....94

Table 6. *Coefficient of Determination of Averaged Semester Grades and Benchmark Raw Score Percents in Literacy*.....95

Table 7. *Coefficient of Determination of Averaged Semester Grades and Benchmark Raw Score Percents in Math*.....96

Table A1. *Survey Results*.....120

Table A2. *Survey Results - Elementary*.....122

Table A3. *Survey Results - Intermediate*.....124

Table A4. *Survey Results - Middle School*.....126

Table A5. *Survey Results - High School*.....128

## KEY TO ABBREVIATIONS

AIP	Academic Improvement Plan
AYP	Adequate Yearly Progress
ACTAAP	Arkansas Comprehensive Testing and Accountability Program
ADE	Arkansas Department of Education
$r^2$	Coefficient of Determination
CRT	Criterion-Referenced Test
EOC	End-of-Course Exams
ITBS	Iowa Test of Basic Skills
NCLB	No Child Left Behind
NRT	Norm-Referenced Test
$r$	Pearson $r$ Correlation Coefficient

## CHAPTER ONE - INTRODUCTION

### *Background*

*In a perfect world, there would be no grades. Students and teachers would work together until students reached a satisfactory level of achievement of intended knowledge and skills. (Guskey, 2009, p. 67)*

There has been a strong consensus among educational experts that knowledge about assessment for teachers is fundamental to effective teaching and such knowledge is vital to student achievement (Trumbull & Farr, 2000). This is especially true in the current educational reform climate. With the passage of the No Child Left Behind Act (NCLB) the emphasis on assessment was pronounced (United States Department of Education [DOE], 2001). The law called for states to create and implement criterion-referenced tests (Arkansas Department of Education [ADE], 2008b). It became necessary for states to target specific learning goals and then test those objectives (Carter, 2007). Even with the failed reauthorization of NCLB in 2007, there is still a concentration on all things standards-based. As a result, the next expected continuation of this process is

to develop new curricula and adopt new grading practices (Kennedy-Manzo, 2008).

However this is one primary area in teacher education programs that has been lacking (Guskey, 2009). When considering the emphasis grades play in public education;

It seems unthinkable that so little attention has been paid to establishing systems that are soundly based on good measurement principles. One might also find it inexplicable that teachers have been so ill prepared to apply such measurement principles in their classrooms. (Trumbull & Farr, 2000, p. 4)

This lack of preparation results in grades that are difficult to defend and has become more evident with the increased focus on assessment (Guskey). As a result, grading practices should be included as part of the reform process because traditional grading is insufficient in assessing student learning, growth, and development (Barnwell, 2008).

Where do the sources of teachers' grading practices originate? First and foremost, the policies and practices teachers experienced as students are the primary source because it is human nature to follow the path that is based on individual prior experience. Second, teachers base practices on personal ideas of teaching and learning.

Third, district, building, department, or grade-level policies on grading and reporting have to be considered. Last, what teachers learned about grading and reporting in undergraduate teacher preparation programs also has an impact on practice (Guskey, 2009; Guskey & Bailey, 2001).

If the goal, as required by the original NCLB legislation, was to meet Adequate Yearly Progress (AYP) by ensuring that a significant number of students are proficient or advanced (DOE, 2001), the grading systems in place must also reflect those same achievement levels. In Arkansas, the rules governing the Arkansas Comprehensive Testing, Assessment and Accountability Program (ACTAAP) handed down by the ADE; the criteria for meeting the mandated AYP is spelled out in detail (ADE, 2007).

Despite the detailed instructions supplied by the ADE regarding the fulfillment of NCLB, very little guidance has been provided to districts pertaining to grades. In ADE's (2007) standards for accreditation report, the only reference to grading procedures stated,

Grades assigned to students for performance in a course shall reflect only the extent to which a student has achieved the expressed academic objectives of the course. Grades that are aligned with other educational objectives such as the student

learning expectations contained in the curriculum frameworks may also be given. (p. 56)

An important starting point for the development of a grading system is the consideration of the purpose of grading students (Marzano, 2000; Trumbull & Farr, 2000). The wide variation in traditional grading practices is due in part to the lack of clarification of purpose. For those who see the purpose as helping students to master certain knowledge and skills, the status of a student's achievement against an explicit standard is important. For those who see the purpose as developmental, grades describe the effort and progress the students are making (Guskey & Bailey, 2001).

Marzano (2000) in his book, *Transforming Classroom Grading*, discussed four factors that teachers commonly include in grading: academic achievement, effort, behavior, and guidance. When teachers use academic achievement as a grading criterion, they assign grades in a manner relative to the amount of content students learn (Marzano, 2000). If they learn a great deal of content, the grade is high; if they learn very little content, the grade is low. However when effort is a sign, students who try harder receive a higher grade than those whose achievement was at the same level but have put less effort into the work. Furthermore,

Marzano discussed that behavior is often incorporated into grading practices and is interpreted as the extent to which students followed classroom rules and procedures. Lastly, he believed, attendance is most commonly used to lower grades: "Perfect attendance and punctuality do not increase a student's grades" (2000, p. 29).

### *Theoretical Underpinnings*

An ongoing philosophical discussion centers on questions such as why do teachers grade and furthermore should they? Schools grade to provide feedback in the following instances: inform parents, account to community, recognize good work, identify unacceptable work, promote student self-evaluation, and identify instructional gaps. Schools use grades to motivate by encouraging students to improve or keep working and rewarding students who are doing well (Guskey & Bailey, 2001; Huhn, 2005; Trumbull & Farr, 2000). Wright and Wise (1988) found that academic achievement and effort considered together account for about eighty percent of what differentiates one grade from another. Grades also help to sort students by making placement or grouping decisions (Guskey, 2008). They also certify competence, permit graduation, advance students to next grade, and predict future achievement.



How do grades and federal mandated assessments make a connection? In essence, what can a set of test scores tell about the quality of education and its relation to student performance? No matter the philosophical stance, it has become apparent that one set of test scores only provide a snapshot of student achievement (Marzano, 2006). Grades earned over an extended period of time on the other hand will offer a more complete picture of student learning (Scriffiny, 2008).

*Statement of the Problem*

A virtual revolution in assessment practices due in part to NCLB is moving at a fast-pace; however, the evolution of grading practices has been much slower. Sometimes current grading systems subvert the good intentions of reformed assessment systems (Winger, 2005). It is crucial that grades reflect the achievement levels the criterion-referenced tests have required under the law.

In a desire to predetermine student proficiency and achievement levels, schools have been concentrating on grading systems to accurately reflect student progress and how accurately they meet state standards throughout the year (Carter, 2007). While state tests are important communicators of student achievement and allow schools to reform curriculum and instruction long-term, ongoing

information that schools require to incrementally improve instructional programs has not been made available (Herman & Baker, 2005). Furthermore, the learning problems of students with the most need have not been addressed (Guskey, 2008).

However, a new wealth of immediate student data presents educators with decision making information. It permits stakeholders time to consider program decisions and evaluate teacher effectiveness (Marsh, Payne, & Hamilton, 2006).

With the use of accurate measures and timely access to the analysis of school/district progress, schools now can determine the amount and nature of academic growth that each student needs and then organize themselves to accomplish these learning goals.

(Olson, 2007, p. 11)

A downside to all of this measuring of student progress has been that purchased off the shelf tests are used and not always linked to state standards (Baker, Linn, & Herman, 2002). A concern has been that educators will focus on the test rather than the standards (Linn, 1998). However, Arkansas has been commended for the ability to link frameworks to national standards (ADE, 2007). For example, the National Council on Teaching Mathematics

(NCTM) recognized Arkansas for its standards being closely aligned to those at the national level (ADE, 2007). Since Arkansas requires its state mandated assessments to be closely aligned with state standards (ADE, 2004), it is not surprising that the natural progression would be to align grading practices with state standards.

#### *Purpose of the Study*

The rationale for the study is to help Arkansas districts achieve AYP through improved grading practices as state funding is directly linked to test scores (ADE, 2004). Districts accurately predicting student achievement through grading can target student weakness prior to benchmark testing and focus efforts on direct remediation. If target areas are identifiable, districts can restructure curricula more effectively and efficiently (Carter, 2007).

One way to ensure that accurate decisions are made at the district and school level is to review the grading practices in place. If classroom grades are a true indicator of student achievement, there should be a correlation between the semester grades and the proficiency rating on the criterion-referenced tests. As a result, it is essential to find an accurate predictor of student achievement on the Arkansas Benchmark Test by examining grading practices and establishing whether a district can

use them to determine achievement on these state mandated assessments.

### *Questions*

There were several questions addressed in this study to conclusively answer the hypotheses.

1. What relationship exists between the first and second semester grades in three consecutive school years including 2006, 2007, and 2008 in the sixth, seventh, and eighth grades?
2. What relationship exists between semester grades and the spring benchmark examination in three consecutive school years including 2006, 2007, and 2008 in the sixth, seventh, and eighth grades?
3. What were educators' attitudes concerning grading practices and the relationship to student achievement on the Arkansas Benchmark Test?

### *Independent Variable*

The independent variable in the study was the averaged semester grades in math and literacy for the sixth, seventh, and eighth grades during the 2006, 2007, and 2008 school years.

*Dependent Variable*

The dependent variable in the study was the raw score percents on the Arkansas Benchmark Test in math and literacy for the sixth, seventh, and eighth grades during the 2006, 2007, and 2008 school years.

*Hypotheses**Null Hypothesis*

The semester grades in literacy and math in a Northwest Arkansas Middle School was not an accurate predictor of student achievement on the Arkansas Benchmark Test.

*Alternate Hypothesis*

The semester grades in literacy and math in a Northwest Arkansas Middle school was an accurate predictor of student achievement on the Arkansas Benchmark Test.

*Limitations of the Study*

*Extraneous factors.* There were important outside factors to semester grades and student achievement on the state assessment such as teacher quality, curriculum quality, parental involvement, socio-economic status, and language barriers. These were impossible to measure within the constraints of the study. Efforts were made to reduce the impact of these issues by limiting the sample group to

only students who participated in both semesters and the Arkansas Benchmark test.

*Research design.* The study was confined to one Northwest Arkansas Middle School.

*Survey.* The survey was designed by the researcher and it was assumed that all respondents answered honestly.

*Test design.* A potential imitation included the criterion-referenced benchmark test. The degree of difficulty changes from year to year as does the cut score which identifies proficiency.

#### *Definition of Terms*

*Adequate Yearly Progress (AYP).* An individual state's measure of yearly progress toward achieving state academic standards, as described in the NCLB legislation. AYP is the minimum level of improvement that states, school districts, and schools must achieve each year (Ravitch, 2007).

*Arkansas Comprehensive Testing, Assessment and Accountability Program (ACTAAP).* A comprehensive system that concentrates on high academic standards, professional development, student assessments, and accountability for all schools. The ACTAAP is also referred to as the Arkansas Benchmark (ADE, 2004).

*Bubble kids.* Students whose current levels of achievement place them near the state's cutoff for determining proficiency (Ravitch, 2007).

*Criterion-Referenced Tests (CRT).* An assessment that measures a student's mastery of skills or concepts set forth in a list of criteria, typically a set of performance objectives or standards. Such tests are designed to measure how thoroughly a student has learned a particular body of knowledge without regard to how well other students have learned it (Ravitch, 2007).

*Formal assessment.* An assessment that collects data using a standardized test in a standardized testing environment (Ravitch, 2007).

*Formative assessment.* Any assessment used by educators to evaluate students' knowledge and understanding of particular content and then to adjust and plan further instructional practices accordingly to improve student achievement in that area (Ravitch, 2007).

*Grade.* A judgment on student performance or conduct, rendered usually either as a letter from *A* to *F* (with *A* representing excellence and *F* representing failure) or as a number, generally from 0 to 100 representing a perfect performance. Teachers may award grades for test

performance, classroom participation, homework, or other student work (Ravitch, 2007).

*Informal assessment.* An assessment that collects data by anything other than a standardized test (Ravitch, 2007).

*No Child Left Behind Act.* A legislative act initiated by the Bush Administration to establish accountability for the nation's public schools through a measurement of Adequate Yearly Progress. Schools and districts are supposed to achieve a goal of 100 percent proficiency in reading and mathematics for every subgroup by the 2013-2014 school year (DOE, 2001).

*School Improvement.* A term used to designate an Arkansas school district which does not meet Adequate Yearly Progress (ADE, 2004).

*Standard.* An officially sanctioned description of what a student is expected to learn and how well it should be learned in specific subjects taught in school. Standards may be created by school districts, states, federal agencies, subject matter organizations, or advocacy groups (Ravitch, 2007).

*Standards-based grade system.* A grading system which measures student progress against a set of fixed standards (Ravitch, 2007).



*Student Achievement.* A definitive measure of a student's academic growth through norm-referenced and criterion-referenced test batteries (Ravitch, 2007).

*Summative assessment.* An assessment used to document students' achievement at the end of a unit or course or an evaluation of the end product of students' learning activity (Ravitch, 2007).

### *Summary*

The federal NCLB mandate required each state to develop a criterion-referenced test and to establish AYP. To successfully meet this law, there must also be a link between the curriculum taught in the schools and the ensuing grading practices. Since funding has been tied to AYP, public schools are examining every available option in order to meet these mandated goals. An examination of school gradebooks as a potential predictor of achievement, and a source for remediation of weaknesses of individual students is imperative. Whether or not these grading practices are aligned with student achievement on the Arkansas Benchmark Test is worthy of investigation.

In chapter two a review of the literature surrounding the theory and research behind grading, assessments, accountability and testing was provided. In chapter three the research design was discussed and in four the data from

this research was analyzed. Furthermore in chapter five an implication for schools and a recommendation for continued study was presented.

## CHAPTER TWO - LITERATURE REVIEW

### *Background*

In recent years, public education has spent a great deal of time, money, and energy attempting to improve procedures for grading and reporting student learning. Educators have recognized the inadequacies in current policies and practices and have been convinced of the need for change (Guskey, 2009; O'Connor, 2007). This was in part due to the huge gap between grading practices and the federally mandated assessments brought forth under No Child Left Behind (NCLB) (Clymer & William, 2007; DOE, 2001).

This chapter examined the literature surrounding current grading practices and the ensuing assessments available to educational practitioners. It would have been incomplete without an examination of the NCLB mandate and the effect on public school accountability as well as an inspection of the different assessment categories and the function of standardized testing.

### *History*

Starting with the historical practices of grading, Trumbull and Farr (2000) offered a thorough examination of the history surrounding grading practices in education.

While not current, the research was certainly worth examination. Trumbull and Farr documented that in the early 1900s, elementary teachers used written descriptions to document student learning and high school teachers introduced percentages as a way to certify students' accomplishments in subject areas.

According to the authors, in the early nineteen twenties, teachers turned to grading scales with fewer and larger categories, such as excellent, average, and poor. Within another decade grading on the curve became increasingly popular as educators sought to minimize the subjective nature of scoring (Trumbull & Farr, 2000). Marzano (2006) recognized that, "arguably the most well entrenched tradition in U.S. education is the overall grade" (p. 105).

It was not surprising that grading cycles have come full circle as educators seek reform. In current grading practices, schools turn back to descriptive terms which encompass a large range of ability (Scriffiny, 2008). However, one difference has been the language currently being adopted for assessment and the fact that it is in line with the terminology stated in the NCLB law. Descriptors such as basic and proficient have been now found in school reporting systems (DOE, 2001) or at least

being championed by experts in the field (Guskey, 2009; Marzano, 2006).

Like grading practices, assessment policy has changed over time. However, unlike today, early assessments were not dictated by government sanction. They were informal, teacher-made tests, which were not lacking in depth (Crone, 2004). The development of the Stanford Achievement Test in 1923 allowed for standardized testing and opened the door for this type of use which increased over time. Between 1941 and 1960, these formal assessments held students and curriculums accountable not public schools (Crone). NCLB now utilizes both criterion-referenced and norm-referenced tests to hold schools accountable by government sanction (DOE, 2001).

The 1965 Elementary and Secondary Education Act provided Title funds to help educate low-income students and testing became the means to judge the program's effectiveness (Crone, 2004; Guifoyle, 2006). One test used to evaluate the success of the Elementary and Secondary Education Act was the National Assessment of Educational Progress (NAEP) developed in the 1960s by the Education Commission of the States. It is administered to nine, thirteen, and seventeen year olds in math and literacy and was designed to measure progress (Crone). Its current

application assists in the diagnosis of a state's testing programs.

A 1983 report by the National Commission on Excellence in Education spotlighted nation-wide attention on public schools. *A Nation at Risk: The Imperative for Education Reform*, stated the national education system was in complete disarray and such was the status of education that it compromised the country's preeminence, technologically and militarily (Wong, & Nicotera, 2007). It was not until much later that NCLB legislation put into action a federal accountability system. The mandate signed into law in 2002 emphasized high stakes testing (DOE, 2001).

NCLB allowed states to create achievement tests (DOE, 2001), but how does the federal government ensure a real measure of student achievement has been accomplished? The NAEP test is administered to a sample of fourth and eighth graders from each state every other year as a means to present a comparison baseline. States whose students scored well on state mandated tests but poorly on the NAEP will be subject to examination (Cavanaugh, 2007). Because NAEP is the only standardized test administered to a representative sample of students across the nation, it has been often referred to as the *Nation's Report Card* (italics added). Since 1969, assessments have been conducted periodically in

reading, mathematics, science, history, geography, writing and other fields to determine what students know and can do in those subject areas (Crone, 2004; Guilfoyle, 2006). NAEP results are reported both as scores and also as performance levels (Cavanaugh, 2008). The names were similar to those used to report Arkansas' benchmarks, though they represented slightly different groupings of students (ADE, 2004).

Although data from state testing programs have shown increasing proportions of students reaching or surpassing the proficiency bar, some experts have questioned the validity of such gains. Those results have raised eyebrows, in part, because trend lines are rising much more rapidly on state-developed tests than on the NAEP (Pollard, 2008). To explore this issue, the EPE Research Center performed an analysis comparing trends on NAEP and state-developed assessments between 2003 and 2007. Data were available for 42 states. In 16 states, gains in the percent of students reaching proficiency in 8th grade math were at least ten percentage points greater on state-developed assessments (Pollard). Overall, the report stated about 80 percent of states experienced a faster growth rate on state tests, while only eight states had larger gains on NAEP than on state assessments (Pollard).

NCLB mandated all districts reach one-hundred percent proficiency of student achievement on state-mandated tests by the end of the 2013-2014 school year (DOE, 2001). As well as designing achievement tests, states are responsible for the following: defining the standards for which students are accountable, classifying proficiency levels, and setting cut points across the distribution of scale scores (ADE, 2004). As a result, these indicators varied drastically among the different states (Fuller, Wright, Gesicki, & Kang, 2007).

States developed and administered tests specifying what constituted an allowable proficiency rating for each grade. This variety permissible in the legislation has caused groups such as the National Association of Secondary School Principals (NASPP) to voice concerns. In a position statement, NASPP asked Congress to create an independent panel of researchers and educators to develop common guideline for proficiency in mathematics and literacy (Kennedy-Manzo, 2008). They say, "The irony is that we have 50 states, which have 50 different definitions of proficiency, and NCLB never even describes what is meant by proficiency" (Kennedy-Manzo, 2008, p. 6). Analysts predicted that by the 2013-2014 school year, a majority of



school districts would not meet AYP requirements (Goldschmidt & Choi, 2007).

What does the future hold for NCLB? This legislation saw its fifth anniversary of the federal bi-partisan legislation come and go. Reauthorization for the mandate was due in 2007, and the calls for change have been coming from even those who have typically supported the legislation, especially the conservatives who voted overwhelmingly for the original bill (Wilcox, 2007). The Fordham Institute in Washington D.C. surveyed twenty education insiders; all but one of the respondents believed the legislation would be held up until after the 2008 presidential election, and a majority felt only small adjustments would be made. They also believed the core of any change would center on a growth model plan which would integrate a variety of measures for accountability (Loup & Petrilli, 2005).

According to Wilcox (2007), Weaver, president of the National Education Association, recommended two ways to improve current accountability systems and help to create a more fair and workable plan. His first suggestion was the use of multiple measures and methods to gauge achievement and school quality to determine school effectiveness. He believed these measures should gauge growth over time and

not be solely based on a certain proficiency level (Wilcox, 2007). States must take every precaution to create accountability systems which avoid unintended, negative results (Stecher & Hamilton, 2002). The goal was to meet federal regulations and use reform measures to actually drive curriculum changes thus benefiting student achievement.

In Arkansas, four factors contributed to a school's Adequate Yearly Progress (AYP) and whether a district was placed on the school improvement list. The first factor is a student assessment in both mathematics and literacy. This is a criterion-reference test aligned to state standards at each grade level three through eight. There is also End of Course exams for Algebra I, Geometry, Algebra II, and Biology as well as an Eleventh Grade Literacy test (ADE, 2005). The second aspect necessary to achieve proficiency is the requirement that ninety-five percent of all eligible students must participate in these academic assessments (ADE, 2004). The third facet is that at least one other additional indicator is necessary; for example, one requirement might be that attendance rates improve by a specified margin each school year (ADE, 2005).

The fourth and final feature is the inclusion of a safe harbor provision. A population makes safe harbor when

it decreases the percent of students performing below proficient by ten percent. In Arkansas, all four indicators hold for the combined population as well as each eligible sub-group. Sub-groups include; economically disadvantaged, racial/ethnic groups, students with disabilities, and Limited English Proficiency. They are considered eligible when the total sub-group population for a building is forty or more students (ADE, 2005).

#### *Purpose of Grading*

"One of the most difficult aspects of standards-based reform is the development of fair, accurate, and defensible procedures for grading and reporting student learning" (Guskey, 2009, p. 57). So the question begs, why grade at all? The simplest and most compelling reason that teachers grade pupils is because of the requirements placed upon teachers to do so. Grading is one kind of official assessment that teachers are required to do (Airasian, 2000; Guskey, 2009).

Mandel (2006) discussed the lack of guidance novice teachers receive in the reason for grading. New teachers want to grade according to school policy but still be fair to students. They want the grades to be accurate but not hurt a student's self-esteem, and they don't want to have to spend hours figuring out grades. Efficient and fair

grading, one of the most fundamental teacher tasks, is not a skill normally taught in education classes or new teacher professional development (Mandel, 2006). However, according to Reeves (2008),

The difference between failure and the honor roll often depends on the grading policies of the teacher. To reduce the failure rate, schools don't need a new curriculum, a new principal, new teachers, or new technology. They just need a better grading system.

(¶ 1)

The first purpose for grading was that teachers needed to communicate the achievement status of students to parents and others. Grading and reporting provided parents and other interested persons with information about the child's progress in school (Davis, 1999; Guskey & Bailey, 2001; Marzano, 2000; Trumbull & Farr, 2000; Winger, 2005). To some extent, it also served to involve parents in educational processes. Marzano's (2000) belief was that grades provided feedback about student achievement, and he reiterated this purpose has been highly valued by both teachers and students. Grades should serve to inform what students know and understand, provide sufficient understanding, and offer a grade that accurately reflects this (Huhn, 2005).

The second purpose was grades are issued to provide information students can use for self-evaluation. Grading and reporting offered students information about the level or adequacy of academic achievement and performance in school (Davis, 1999; Guskey & Bailey, 2001; Marzano, 2006; Trumbull & Farr, 2000). Third, grades have been used to select, identify, or group students for certain educational paths or programs. Grades have been a primary source of information used to select students for special programs. High grades are typically required for entry into gifted education programs (Davis; Guskey & Bailey). It is important to know if grading practices are an accurate reflection of student achievement if grades are being used for placement purposes. Alternately, low grades are often the first indicator of learning problems that result in students' placement in special needs programs. Grades have been also used as a criterion for entry into colleges and universities (Guskey & Bailey; Trumbull & Farr).

The fourth purpose of grades was to provide incentives for students to learn. Although some may debate the idea, extensive evidence shows that grades and other reporting methods are important factors in determining the amount of effort that students put forth and how seriously they regard any learning or assessment task (Davis, 1999; Guskey

& Bailey, 2001). Marzano (2000) referred to this as motivation, and he also indicated some educators strongly object to this use. Kohn (1989; 2003) objected to the use of grades as a tool for rewards.

Guskey and Bailey (2001) identified grades as a way to evaluate the effectiveness of an instructional program. Grades are used by teachers to make initial decisions about student strengths and weakness in order to group them for instruction. Comparisons of grade distributions and other reporting evidence are frequently used to judge the value or effectiveness of new programs and instructional techniques (Trumbull & Farr, 2000).

A final purpose of grades was as evidence of students' lack of effort or inappropriate responsibility. Grades and other reporting devices are frequently used to document unsuitable behaviors on the part of certain students, and some teachers threaten students with poor grades in an effort to coerce more acceptable behaviors (Guskey & Bailey, 2001).

### *Theory*

It has become important that grading components align with the state and district standards; some may be drawn primarily from content or skills already identified by such standards (Scriffiny, 2008; Trumbull & Farr, 2000). A grade

that has been separated into distinct components on the basis of key learning becomes a meaningful communication, both to students and parents alike, about what students have and have not mastered (Winger, 2005).

So then what role do standards play between grading and standardized assessment? The public demanded reform and NCLB was a bi-partisan measure, so it is not likely to go away (Hoff, 2008; Wilcox, 2007). That is not to say that NCLB has no critics. One criticism was the law did not require deep-lasting reforms to take place and only measures growth against fixed standards (Elmore, 2003). Kohn (2001) believed the standards are causing the destruction of teacher innovation by creating a teacher proof curriculum and destroying a district's creativity in a desire to meet AYP.

If the mandated tests are criterion-referenced and the common-sense link to grading practices is that that they also must be tied to standards, the solution should be simple. Unfortunately the standards themselves have been cause for confusion. There are seven different types of standards that have surfaced since the movement began in the early 1900s. As defined by Trumbull and Farr (2000), they are as follows:

1. Content standards: What should students know and be able to do?
2. Performance standards: How good is good enough?
3. Delivery standards: What materials and resources are necessary to achieve established performance standards?
4. Opportunity standards: What kinds of instruction are necessary to achieve established performance standards?
5. Instructional standards: What constitutes exemplary instructional practice?
6. Assessment standards: How do we evaluate the quality and validity of our assessment tools?
7. Process standards: What guidelines should we follow for developing and implementing standards? (p. 158)

The current emphasis on established content standards has focused teaching on designated knowledge and skills. To avoid the danger of viewing the standards and benchmarks as the only content to cover, educators should frame the standards and benchmarks in terms of desired performances and ensure that the performances are as authentic as possible (McTighe & O'Connor, 2005).

Not all education researchers have been supportive of the content standards movement. Popham (2006a) believed



that most states have far too many content standards: "Moreover, they are poorly conceptualized either for teaching or testing" (p. 87). Furthermore, he believed that in some states, content standards have been little more than category labels describing collections of curricular aims with no real connection. When teachers and test makers become overwhelmed by too many standards, any test-based accountability program is certain to stumble. "The proliferation of standards developed at the national and state levels turns the preparation of a meaningful classroom curriculum into a daunting task" (Harris & Carr, 1996 p. 1). When test makers are unable to assess all of the state's sprawling curricular aims, test designers settle for a sampling (Marzano, 2006; Popham, 2006a). This makes it difficult for teachers to provide targeted instruction if they do not know which of the specific curricular aims a student has or has not mastered (Marzano, 2006).

It is difficult to link grades to standardized assessment, to know exactly which instructional practices and or grading practices directly relate to student achievement on the mandated tests (Carter, 2007). Quality of instruction has not been directly measured in many accountability systems because few assessment tools have

the potential to directly measure the quality of classroom practice on a large-scale basis (Junker, Weisburg, Matsumura, Crosson, Wolf, Levison, & Resnick, 2006).

However, researchers have been seeking data to support powerful teaching and learning environments which aid in student achievement on standardized assessments. Junker, et al (2006) in a Technical Report from the National Center for Research on Evaluation Standards and Student Testing developed the following theory to support successful teaching and learning environments. The first was a learner-centered method. Teachers were able to recognize predictable misconceptions of students where the mastery of particular subject matter was difficult. The second was referred to as knowledge centered. With this, teachers must teach some subject matter in depth and provide enough examples in which the same concept is at work so that students can grasp the core concepts in an area. The third model provided by the researchers was assessment centered. Teachers must help students develop a clear understanding of what they should know and be able to do, setting learning goals and monitoring progress together. Fourth was community centered; "teachers must arrange classroom activities and help students organize work in ways that promoted the kind of intellectual camaraderie and attitudes

toward learning that build academic community" (Junker, et. al, 2006 p. 3).

As part of RAND's evaluation of the Federal Systemic Initiatives program of the 1990s, Klien, Hamilton, McCaffrey, Stecher, Robyn, and Burroughs (2000) studied instructional practice and student achievement with 627 teachers distributed over three elementary middle grade levels and six sites. They found substantial variation in educational practice within schools, and after controlling for background variables, a generally weak but positive relationship between frequency of reform teaching behaviors and student achievement. The relationship was somewhat stronger when achievement was measured with open-response tests than with multiple choice tests (Klien, Hamilton, McCaffrey, Stecher, Robyn, & Burroughs, 2000).

For a teacher who wants his or her students to learn big ideas and gain long-term understanding, assessment means being keenly aware of what students know and understand (Wiggins & McTighe, 2005), having sufficient evidence of this understanding, and offering a grade that accurately reflects this.

Teachers often lament their students' myopic focus on grades. Frustration mounts when students ask 'How many points is this worth?' I don't believe the

problem lies with losing track of what grading and assessment are supposed to mean. Students ask for extra-credit just prior to progress reports etc. A grade of seventy-five percent should mean a student knows three-fourths of the material. (Huhn, 2005, p. 84)

The challenge of effective grading has been daunting indeed. Even in the hands of highly qualified, well trained, sophisticated teachers with a well structured curriculum, quality assessment tools must be used in quality ways to make a difference. A CRESST Report by Herman, Osmundson, Ayaly, Schneider, and Timms (2006) discussed case studies that were done by Bell and Cowie in 2001; these studies dealt with teachers' use of assessment to promote student learning. They stated: "Through the assessment of students' needs and the monitoring of student progress, learning sequences can be appropriately designed, instruction adjusted during the course of learning, and programs refined to be more effective in promoting student learning goals" (p. 1).

A great deal of research has been completed on accountability systems. Stapleman, (2000) in a McRel Policy Brief, examined one such study which presented six points to consider when developing an accountability system.

First, standards-based systems improve learning when all components work together. Second, assessments must be aligned with content standards in order for the assessment to be fair and accurate. It is unfair to mandate educators to teach a certain set of content standards but administer an accountability test which covers something else entirely. Third, there must be high-stakes consequences attached in order to motivate schools to improve performance. The Brief pointed out that in this litigious society, the accuracy of these high-stakes consequences will be challenged. Fourth, the accountability system should provide several indicators and not hinge on a single test score. Possible variables included student achievement, attendance, drop-out rates, and graduation rates. This point was a common theme among the various studies developed on accountability systems. Fifth, there needs to be an assistance measure in place to help struggling schools. Sixth and lastly, the report showed that a strong system of rewards and sanctions must be legislated to afford the strength in the mandate to maintain the necessary compliance by the districts. The report also indicated that there was little evidence to support that these rewards or sanctions actually work (Stapleman, 2000).

Another model that emerged from a series of experts centering on a standards-based, state-level accountability system contained similar components found in the McRel Policy Brief. This model also called for an alignment of standards and assessments (Sanders & Horn, 1995). Kohn (2001) provided criteria for judging standards. He believes standards should be non-specific. The more specific the standard, the further students and teachers are distanced from the learning process. There is no room for creativity and investigation when the goal simply is to cover massive amounts of material. He doesn't believe that standards have to be measurable, and he stated, "Measurable outcomes may be the least significant results of learning" (Kohn, 2001, ¶ 3). Kohn also has a problem with uniform standards where all students must learn exactly the same thing, and lastly, he wanted standards to be considered guidelines rather than mandates.

The second part of the model for standards-based accountability systems developed by Sanders and Horn (1995), like the McRel Brief, consisted of a rating system for school performance which contained multiple indicators such as student achievement, attendance, drop-out rates, and graduation rates. It also similarly considered assistance to struggling schools as well as a system for

rewards and sanctions. This study differed from the previous report as it included a method for reporting performance.

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) also developed criteria for an accountability system. Like the two previous reports, Baker, Linn, and Herman (2002) also placed an emphasis on employing different types of data from multiple sources. Furthermore, it called for a report card where results should be made available and understandable with all elements in the system explicitly identified. A difference in this report from others was that it took into account the performance of all students including subgroups that historically have been difficult to assess. Also, rules for determining adequate progress of schools and individuals must avoid wrongful conclusions that are actually attributable to measurement errors in test results (Baker, Linn, & Herman, 2002).

As earlier stated, with the advent of criterion-referenced tests and the development of state-developed content standards, the next natural progression was to realign grading practices to the standards (Guskey & Bailey, 2001). Just as it sounds, this practice involved measuring student proficiency on well-defined course

objectives. Standards-based grading ensured students are being graded over the material they are being held accountable for (Guskey, 2009). Grades should have meaning. An A means that the student has completed proficient work on all course objectives and advanced work on some objectives. This type of grading provides systematic and extensive feedback on assignments and helps to send students the message that they can and should do homework as practice (Scriffiny, 2008).

Furthermore, this type of grading practice also helps the classroom teacher as it reduced paperwork. Christopher (2008) stated, "I don't assess student mastery of any objective until I am confident that a reasonable number of students will score proficiently, and that makes each assessment mean much more" (p. 74). Once the paperwork is done, standards-based grading helps teachers to adjust instruction. The standards-based gradebook provides a wealth of information. In a traditional gradebook, a student would assume they are doing okay, but in the standards-based gradebook it reveals a crucial concept was not grasped (Scriffany, 2008). Projects are graded to the standards without a percentage grade. A grade is given for each standard being assessed so that one test or project often has several different grades, each indicating



progress toward a different standard. This provides more meaningful feedback for students and parents (Christopher, 2008).

In this type of grading practice, students who struggle can continue to retest and use alternate assessments until they show proficiency, and they are not penalized for needing extended time (McTighe & O'Connor, 2005; Guskey, 2009). Students are not permitted to submit substandard work without being asked to revise. A critical part of the standards-based gradebook is the performance assessment. Performance assessments yield evidence that reveals understanding. It requires students to transfer knowledge (McTighe & O'Connor, 2005).

Teachers should set up realistic, authentic contexts for assessment that enable students to apply their learning thoughtfully and flexibly, thereby demonstrating their understanding of the content standards. (p. 12)

According to McTighe and O'Connor (2005), performance assessments are typically open-ended and do not yield a single, correct answer or solution process. Also, a rubric is a widely used evaluation tool consisting of criteria, a measurement scale, and descriptions of the characteristic for each score point. Well-developed rubrics communicate

the important dimensions or elements of quality in a product or performance and guide educators in evaluating student work. Classroom assessments and grading practices should focus on how well the student mastered the designated knowledge and skill not on when (McTighe & O'Connor, 2005). There are some teachers who believe that students work habits, responsibility and attitudes are also important, but it is important to report academic and nonacademic factors separately (Winger, 2005)

#### *Reform*

As educators, one thing which can be controlled is the grading practice and the way teachers assess students. There are so many things outside of a teacher's control, for example, socio-economic level, home life, class size, parents (Kohn, 2004). However, assessment can be controlled. It's a launch-pad to other reforms. Education reform has caused a revision of curriculum to identify clear and concise standards and precise levels of mastery evidenced through assessment (Scriffiny, 2008). This reform over the last half century has been placed squarely on the shoulders of accountability and assessment, and since the 21<sup>st</sup> century, it has been known by the name No Child Left Behind.

While testing and assessment have both critics and proponents, there are several reasons for the appeal of mandated assessment with all of the players, i.e. the public, policymakers, and educators as agents of reform. One of the first and primary reasons for the popularity of assessment as a gauge of a reform's success or failure was that it is fairly inexpensive compared to other measures. Even at an approximately 517 million dollar price tag, it is a small portion of a 500 billion dollar budget spent annually by the United States Department of Education (Toch, 2006). Expensive items in lieu of assessment measures involve hiring more certified staff or increasing instruction time and reducing class size. In difficult budget times, it is unlikely that these tools will be implemented as resources and are already stretched thin even when educators argue the merits (Norton, 2009). All other things being equal, assessment is cheap.

A second reason for the appeal of testing and assessment as a reform tool is that policymakers are able to mandate targets. The original philosophical idea, rightly or wrongly, was that an objective target score is a fair gauge of whether or not reform is successful within a district. Adequate Yearly Progress provides districts with target scores they must meet each year (ADE, 2004). It was

more subjective to require longer lasting, deeper instructional changes inside a classroom and far more expensive (Zellmer, Frontier, & Pheifer, 2006). Furthermore, testing and assessment on the surface was a quick fix for reform. This made it popular with Congress because these requirements are visible within an elected official's term in office. This may have something to do with why NCLB received bi-partisan support and also why reauthorization did not happen until after the 2008 election (Klein, 2008).

Lastly, assessment is appealing because results are easily reported to parents, the public, and the press. Public Agenda has surveyed the public and focus groups, and while the word accountability was seldom used, "Generally, people believe in motivating students, teachers, and administrators to do their best. They also believe in imposing consequences for lack of effort, repeated failure, or demonstrated incompetence" (Johnson, 2003, p. 36).

Testing and assessment have a variety of designs and forms, so first and foremost, the goal of the assessment must factor into the particular choice of assessment. Its design must supply information which meets those desired goals and ensure these records will affect an alteration in the system designed to enhance student achievement (Linn,

1998; Marzano, 2006; Popham, 2006b). Unfortunately, one potential problem has been that there are often conflicting goals between local and state educators. Policymakers must concentrate on the lowest performing schools and meet those requirements first. If blanket policies are implemented statewide, higher-performing schools will be reluctant to move away from programs that are already effective even though student achievement may not yet be maximized (Elmore, 2003; Lewis, 2000). Furthermore, a state may not want to employ strict guidelines while an individual school may want to use these stricter guidelines to force changes within their district (Lewis, 2000), some of which may be for local political reasons rather than those which are educationally sound.

Those responsible for mandating and overseeing assessment reforms must be aware as to what tests actually accomplish. It is crucial to apply these assessments in the manner for which they were designed especially if they are to be part of a legislated accountability system. "With assessment, purpose is everything" (Stiggins, 2008, p. 3). Furthermore, it is essential that educators be consistent with the instructions of the test maker. Using a test for less than its intended purpose will cause the results to be invalid. Critics of the current system believed that,

using fully adaptive assessments would, at long last, enable states to turn the No Child Left Behind law's blunt-force, pass-fail results into much more nuanced relevant and timely information that teachers could use to improve their instruction." (Sokola, Weinberg, Andrzejewski, & Doorey, 2008, p. 27)

There have been four factors to consider when choosing the assessment. First is the test type. Is it an achievement or aptitude test? Achievement and aptitude tests, while similar, measure two different concepts. Achievement tests measure the specific content a student has learned, whereas aptitude tests attempt to predict a student's future behavior or achievement (Laitsch, 2005). Second, for what is the test going to be used for? Is it used for diagnostic purposes, placement purposes, formative evaluation, or summative evaluation? Third, what is the scoring reference that will be used? Are the test scores going to be reported as raw or scale scores? Is this a norm-referenced test or a criterion-referenced test (Bond, 1996; Laitsch, 2005)?

Fourth, not only is the type of assessment key, but the value of the assessment is equally critical as well (Stiggins, 2008). Popham (2003), emeritus professor of education at UCLA, provided three gauges as to whether an

assessment has value. He referred to this as being instructionally sensitive. His definition of instructionally sensitive means it is a test that determines the presence of instructional improvement. The first indicator is the degree of difficulty of the content standards being measured. The second meter is the description of the test's assessed content standards, and the third gauge is the reporting procedures used for group and individual student reports (Popham, 2003).

There have been traditionally two views about the evaluative concepts of assessment. The first, assessment for learning is diagnostic or prescriptive in nature. It is a determinant for placement, instructional planning, or for grouping (Chappuis & Chappuis, 2008; Popham, 2007; Stiggins, 2008).

Also in this view is a measurement for instructional planning decisions which help to clarify and specify how and where a student is taught, or to identify if a student has mastered a set of subskills needed to move on to more difficult curricular aim. These tests are used to help teachers and administrators plan educational programs (Popham, 2006a). According to Chappuis and Chappuis (2008), assessment for learning should help to answer three questions for students: One, where am I going? Two, where

am I now? And, three, how can I close the gap? Feedback is the key because with this type of assessment there is still time to take action, and it should impart a route for students to take to get to where they need to be.

Assessment for learning is "designed to increase, not merely monitor, student confidence, motivation, and learning" (Stiggins, 2008, p. 9).

Assessment of learning is where students demonstrate knowledge of a particular curricular area for progress monitoring or grading purposes. It is evaluative in nature and is used for accountability, rewards, and sanctions. These assessments support student progress decisions (McTighe & O'Connor, 2005; Stiggins, 2008). Since achievement is what has been learned as a result of instruction in schools, only achievement tests measure student progress. A concern brought by educators is that the assessment of learning mandated by NCLB will overshadow assessment for learning as teachers focus on covering materials necessary to achieve AYP (Popham, 2006a).

There are two general categories of assessments to which educators look. The first, informal assessments are the collection of data by anything other than a standardized test. These make up the majority used by the classroom teacher, such as portfolios, teacher observation,



teacher-made tests, and computer-based testing (Rabinowitz, 2001; Tomlinson, 2008). While these are only a few examples, evaluations of this nature impart more accurate diagnostic information since they are not bound by the same constraints as statewide tests (Rabinowitz).

Informal assessments are also made up of three sub-groups: formative, interim or progress, and summative assessment. There is often a great deal of confusion about the roles of these types of assessment. What then is the difference? Formative assessment answers the question "How am I doing?" The data that provides the answer to the question is where the benefits reside (Starkman, 2006). Furthermore, it is how the results are used that separates formative from summative. Formative evaluations are structured assessments designed to gauge the progress of students as measured against specific learning objectives. Such assessments help guide instruction so that teachers and students have a general idea of what learning outcomes are achieved and where further focus is needed. It involves frequent testing, and measurement of student learning is just one component (Chappuis & Chappuis, 2008.)

A more recent assessment alteration is the use of an interim assessment. These are administered periodically throughout the year to monitor student progress at meeting

state standards, usually in math and literacy. These tests provide rapid, regular feedback to students, teachers, and administrators (Marsh, Pane, & Hamilton, 2006; Popham, 2007). One indicator of the importance of interim/progress tests is the rapid increase in availability of such products from commercial test providers (Marsh, Pane, & Hamilton, 2006).

Lastly in the informal sub-group is the summative assessment, which evaluated achievement at the end of specific educational programs. The purpose was to measure the level of student, school, or program success (Chappuis & Chappuis, 2008; Ravitch, 2007). However one problem has been that results from state-mandated tests are often reported in ways that make it difficult for teachers to comprehend, so even if these tests are suggested for use for formative purposes, a lack of teacher comprehension makes this difficult (Chappuis & Chappuis, 2008.)

Tomlinson (2008) puts all of the assessment practices into perspective for the classroom teacher as she distinguishes between assessment of learning, assessment for learning, and assessment as learning:

In many ways, my growth as a teacher slowly and imperfectly followed that progression. I began seeing assessment as judging performance, then as informing

teaching, and finally as informing learning. In reality all those perspectives play a role in effective teaching. The key is where we place the emphasis. (p. 13)

The second category of assessment types is known as formal assessment which is defined as a collection of data using a standardized test in a standardized testing environment (Ravitch, 2007). Due to the magnitude of requirements under NCLB, standardized assessments are the norm for statewide testing purposes. However, to enhance student achievement, the best way is to incorporate a variety of well-rounded student achievement multiple assessment types "because they can combine results from commercially available, standardized tests with those from locally developed, alternative assessments" (Stapleman, 2000, p. 3).

Testing has become big business. It is an unregulated industry whose revenues are skyrocketing. Not only is there a cost in the test itself, but the scoring and reporting of these tests is expensive (Clarke, Madaus, Horn, Ramos, Lynch, & Lynch, 2001; Toch, 2006). Since the results of these high-stakes tests are so important, there is a call to begin regulation (Clarke, et. al). Testing company executives report that states spend \$700 to \$750 million

annually on testing contracts. However, this equates to about one percent of the overall budget. As a result, tests are not examined as closely by the states and local districts as they should be (Toch). Many states do not have the time, finances, or staff to implement tests that align with their state standards. These unaligned tests will give skewed results and lack validity (Toch).

As long as the federal government mandates testing and applies the funding carrot, states have no choice but to struggle daily to comply (Kohn, 2001). In order to validate limited varying resources, i.e. time, money, staff, local districts must employ these tests and the disaggregated data to improve curriculum and instructional practices. Testing is only beneficial if the information gathered is transformed into practices that improve student learning.

A key to the effective use of available resources is to focus and strategically reallocate federal resources...to meet the policy and programmatic issues that are most pressing and that are most likely to improve student achievement. (Cicchinelli, Gaddy, Lefkowitz, & Miller, 2003, p. 3)

It has been difficult to determine a standardized assessment's ability to enhance student learning, but even so, the quality of the assessment is paramount. It is even

more problematic when states adopt the ideology that "test-based accountability systems embody the belief that public education can be improved through a simple strategy" (Strecher & Hamilton, 2002, p. 3). If states and local districts have been spending valuable time and money but not yielding accurate information, precious resources are wasted (Herman & Baker, 2005). For as many different standardized tests available to the consumer, the more varied the ability to assess student knowledge (Popham, 2007).

However, there is good news; often these standardized tests undergo rigorous validation criteria, reliability testing, and standardization procedures from the testing companies (Sanders & Horn, 1995). The rationale underlying reliability is that a test should produce the same score even if the student takes the test on a different day or is administered a version of the test with a different sample of test items (Runyon, Coleman, & Pittenger, 2000). In other words, chance effects should not have a significant influence on test scores. While reliability refers to whether test scores are constant indicators of student performance, validity signifies the degree to which the test items reflect the specified content domain (Runyon, Coleman, & Pittenger, 2000).

There has been a concern that these large-scale external assessments may be unable to measure the academic content and curriculum covered at the local level. Furthermore these tests have drawn criticism from educators and policymakers who believe that they should not be used to make high-stakes decisions because they are limited in ability to measure student attainment of high-quality academic standards (Popham, 2007; Wong & Nicotera, 2007). Educators must be familiar with the way each type of assessment operates to determine the multiple indicators of student performance that will provide enough information to make improvements in instructional practices (Wong & Nicotera). According to Weaver, (2006),

Standards that reflect content mastery alone do not enable accountability and measurement of 21<sup>st</sup> century skills. And without a comprehensive, valid system of measurement, it is impossible to integrate these skills effectively into classroom instruction or monitor whether students have mastered the skills necessary in life and work today. (p. 33)

The Association of American Publishers (AAP 2000) believed standardized tests provide four critically important tasks: First, to identify the instructional requirements of individual students so educators can

respond with effective, targeted teaching and appropriate instructional materials. The second task is to judge students' proficiency in essential basic skills and challenging standards as well as measuring their educational growth over time. Third, standardized tests should help to evaluate the effectiveness of educational programs. And, lastly to monitor schools for educational accountability under NCLB. However, the AAP (2000) cautions those tests should be considered a means to an end and not ends in themselves.

Even within the same category of standardized tests not all components are equal. There are different question types and degrees of difficulty on individual tests. Common formats are items such as multiple-choice. These questions afford an adequate measure for lower level skills such as vocabulary and general principles (Laitsch, 2005). Constructed response offered the best gauge for complex achievement, such as application, inference, and generating hypotheses or conducting experiments. However, test companies are placed under time and money constraints, so often these tests assess only the simplest of skills (Laitsch; Toch, 2006).

Performance and portfolio assessments are not thought to be part of the standardized testing genre but allow for

a demonstration of student competency. In Arkansas, students with special needs, when it is determined the regular test is not appropriate, are allowed to submit a portfolio to show proficiency in math and literacy. These include performance assessments which offer presentations of student work (ADE, 2008a). However, they are extremely time consuming and teachers spend many hours in preparation. Scoring also takes evaluators a number of hours. These assessments are more expensive and difficult to administer, and scores can not be scaled to match regular testing students (Laitsch). Individual states work with test companies to determine a design suitable for these students' needs.

There are two primary types of standardized tests: criterion-referenced tests and norm-referenced tests. Under NCLB (DOE, 2001), states may include either or both of these assessments, and beginning no later than the 2005-2006 school year, a state must administer annual assessments in reading/language arts and math in each of grades three through eight and at least once in grades ten through twelve. Furthermore, beginning no later than the 2007-2008 school year, a state must administer annual assessments in science at least once in grades three



through five, grades six through nine, and grades ten through twelve (DOE, 2001; Guilfoyle, 2006).

Criterion-referenced tests are defined as student knowledge measured against a set of pre-determined standards. Educators choose these tests when they want to determine how well students have mastered a set of skills or a desired curriculum (Ravitch, 2007). Criterion-referenced tests are designed to reflect the knowledge and skills students should know and be able to do in order to display mastery of the academic content (Bond, 1996). In Arkansas, this assessment is required by state statute, rule, or regulation, and is designed by the State to measure student performance/achievement on the State's Academic Content Standards (ADE, 2004).

Cut scores on these criterion-referenced tests developed by the testing company to define proficiency result in an arbitrary number of students scoring above or below the specified number (Laitsch, 2005). The test may be positively or negatively skewed depending on how well the teacher addresses the state mandated content standards. This supports the argument for teaching to the test rather than teaching for student achievement (Laitsch).

Norm-referenced tests are defined as student knowledge measured against other students in their cohort. These

tests measure student performance on a broad range of academic content with test items that differentiate between high and low achievers (Ravitch, 2007). Furthermore, they are chosen to highlight differences in order to rank students. In Arkansas, the norm-referenced assessment is required by state law, rule, or regulation to measure the performance/achievement of Arkansas students relative to the achievement of students nationwide who comprised the norm or standardization group for a particular commercial instrument. This allows students to be compared to peers, but in Arkansas these scores are not factored into AYP (ADE, 2004). On a norm-referenced test, scores are reported so that half of the testers score in the top fifty percent and half in the bottom fifty percent. Items have different degrees of difficulty and those that are too easy or too hard are rejected. These items are not created to match state standards (Laitsch, 2005). In norm-referenced tests, standard scores use the normal curve to report student performance in terms of how many standard deviations the test score is from the mean test score (Laitsch).

Before states choose the type of standardized test, they need to consider three questions. Does the test match the educational goals? Does the test address the content assessed? Does the test provide appropriate interpretations

(Bond, 1996)? Laitsch in his Infobrief (2005) reported that the Association for Supervision and Curriculum Development (ASCD) advocates multiple measures as a gauge for the success of an accountability system. According to Laitsch they also believe in assessments that are

- Fair, balanced, and grounded in the art and science of learning and teaching;
- Reflective of curricular and developmental goals and representative of content those students have had an opportunity to learn;
- Used to inform and improve instruction;
- Designed to accommodate nonnative speakers and special-needs students; and
- Valid, reliable, and supported by professional, scientific, and ethical standards designed to fairly assess the unique and diverse abilities and knowledge base of all students. (¶ 1)

In a desire to predetermine student proficiency and achievement levels, schools are creating or purchasing assessment systems to monitor student progress and how accurately they meet state standards throughout the year (Popham, 2006b). In many states, reporting of annual scores are delivered too late in the year to accurately remediate

student weaknesses (ADE, 2004), so these pre-assessments are essential to raise achievement levels. While the state tests are important communicators of student achievement and allow schools to reform curriculum and instruction long-term, they do not provide ongoing information that schools employ to incrementally improve instructional programs (Herman & Baker, 2005). Furthermore, they do not address the learning problems of students with the most need (Herman & Baker). These state tests are assessments of learning, and districts understand assessments for learning are also a necessity.

Carol Ann Tomlinson (2008) referred to these types of assessments as informative assessments and discussed the ability they have to guide instruction.

I slowly came to realize that the most useful assessment practices would shape how I taught. I began to explore and appreciate two potent principles of informative assessment. First, the greatest power of assessment information lies in its capacity to help me see how to be a better teacher. If I know what students are and are not grasping at a given moment in a sequence of study, I know how to plan our time better. I know when to reteach, when to move

ahead, and when to explain or demonstrate something in another way. Informative assessment is not an end in itself, but the beginning of better instruction.

(p. 11)

Pre-assessments allow for educators to evaluate how students are performing at a single point in time, but if the results are reported immediately and if they are administered at different points throughout the year, growth progress is measurable (McTighe & O'Connor, 2005; Popham, 2007). "Teachers use them to check students' prior knowledge and skill levels, identify student misconceptions, profile learners' interests, and reveal learning-style preferences" (McTighe & O'Connor, p. 12). This affords educators an opportunity not previously available in the public school setting. In order for an accurate measurement to weigh against the annual assessment and to supply accurate instructional opportunities, it is necessary that these assessments be aligned to state mandated content standards, which in turn allows for growth measurement regardless of achievement status (Olson, 2007).

This new wealth of immediate student data presents educators with decision making information. It permits them

to consider program decisions and evaluate teacher effectiveness (Reeves, 2006).

With the use of accurate measures and timely access to the analysis of school/district progress, schools now can determine the amount and nature of academic growth that each student needs and then organize themselves to accomplish these learning goals.

(Olson, 2007, p. 11)

According to Douglas Reeves (2004), CEO and founder of the Center for Performance Assessment, many school districts have started using data to drive decisions to expand student learning and achievement. Schools are learning to use pre-assessments and end of the year test results to evaluate lack of or increases in student achievement. This is a key change because most data-driven decision making in the past was more about looking at end-of-year test results with little or no analysis to tie-in causes. "It was an autopsy. I've never seen a patient get better because of an autopsy" (Pascopella, 2006).

A 2006 Rand study revealed a common set of factors to help explain why some educators tend to use data more and with greater levels of sophistication than others. These included accessibility, quality (real or perceived), motivation, timeliness, staff capacity and support, and

curriculum pacing pressures (Marsh, Pane, & Hamilton, 2006). However, using data-driven decision making does not guarantee effective decision making. The process of translating data into information, knowledge, decisions, and actions is labor intensive (Elmore, 2003), and practitioners need to consider the trade-offs of time spent collecting and analyzing data as well as the costs of providing needed support and infrastructure to facilitate data use (Zellmer, Frontier, & Pheifer, 2006).

When a need is apparent and money is to be made, vendors and service providers have jumped in to fill this gap with a variety of products and services. These are referred to by such names as benchmark tests, progress monitoring systems, and formative assessments (Popham, 2006b). Many of these products are developed to coordinate with state standards and allow schools to administer them regularly, often quarterly, to gauge student progress (Herman & Baker, 2005).

The quality of the assessment is essential: "There is little sense in spending time and money for elaborate testing systems if the tests do not yield accurate, useful information" (Herman & Baker, 2005 p. 50). There are several criteria for determining the validity of these pre-assessment benchmarks. These are as follows: align the

standards and benchmark assessments from the beginning of test development, enhance the diagnostic value through initial item and test structure design, ensure the fairness of benchmark assessments for all students, insist on data showing tests' technical quality, build in utility and hold benchmark testing accountable for meeting its purposes (Herman & Baker, 2005).

Education reform in the 21<sup>st</sup> century has been in the manner of all things standards-based. States set standards, hired testing companies to create norm-referenced and criterion-referenced tests, aligned instructional practices, and embedded formative assessments into classrooms to document the success of student understandings of the standards. The next step was to examine current grading practices and align these techniques to more closely monitor the standards in individual classrooms.

### *Effects*

#### *Grading Practices*

Classroom assessment and grading practices have the potential not only to measure and report learning but also to promote it. Recent research had documented the benefits of regular use of diagnostic and formative assessments



(McTighe & O'Connor, 2005). In opposition to the above thoughts by McTighe and O'Connor, Kohn (2003), a well-known critic of all things standardized, believed there are three main effects of grading practices. Firstly, he believed grades tend to reduce students' interest in the learning itself. The more people are rewarded for doing something, the more they tend to lose interest in whatever they had to do to get the reward. Secondly, grades tend to reduce students' preference for challenging tasks. Students of all ages who have been led to concentrate on getting a good grade are likely to pick the easiest possible assignment if given a choice (Kohn, 2003). The more pressure to get an A, the less inclination to truly challenge oneself. Lastly, grades tend to reduce the quality of students' thinking. Given that students may lose interest in what they're learning as a result of grades, it makes sense that they're also apt to think less deeply (Kohn, 2003).

Opposition to grading by Kohn is not just with standards-based grades; as far back as 1989 Kohn was criticizing the use of grades. He believes grades aren't reliable, valid, or objective. A score on a test is largely a reflection of how the test was written, what skills the teacher decided to assess, what kinds of questions happened

to be left out, and how many points each section was worth. Grades distort the curriculum and encourage the instruction of facts because they are easier to grade (Kohn, 1989). Furthermore, grades waste a lot of time that could be spent on learning. Include all the hours that teachers spend fussing with gradebooks and then factor in the conversations they have with students and parents. Grades encourage cheating, and the more students are led to focus on good grades, the more likely they are to cheat (Kohn, 1989).

According to Guskey and Bailey (2001) and O'Connor (2007), there are several ways to fix potential problems with grading practices. A list of practices teachers should avoid include the following:

- Do not include student behaviors in grades such as effort, participation, and adherence to class rules; instead include only achievement (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not reduce marks on "work" submitted late; instead, provide support for the learner (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not give points for extra credit or use bonus points; seek only evidence that distinctly proves that

more work has resulted in a higher level of achievement (Guskey & Bailey, 2005; O'Connor, 2007).

- Do not punish academic dishonesty with reduced grades; apply other consequences and reassess to determine actual level of achievement (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not consider attendance in grade determination; report absences separately (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not include group scores in grades; use only individual achievement evidence (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not organize information in grading records by assessment methods or simply summarize into a single grade; organize and report evidence by standards/learning goals (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not assign grades using inappropriate or unclear performance standards; provide clear descriptions of achievement expectations (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not assign grades based on student's achievement compared to other students; compare each student's

performance to preset standards (Guskey & Bailey, 2005; O'Connor, 2007).

- Do not rely on evidence gathered using assessments that fail to meet standards of quality; rely only on quality assessments (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not rely only on the mean; consider other measures of central tendency and use professional judgment (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not include zeros in grade determination when evidence is missing or as punishment; use alternatives, such as reassessing to determine real achievement or use "I" for Incomplete or Insufficient Evidence (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not use information from formative assessments and practice to determine grades; use only summative evidence (Guskey & Bailey, 2005; O'Connor, 2007).
- Do not summarize evidence accumulated over time when learning is developmental and will grow with time and repeated opportunities; in those instances, emphasize more recent achievement (Guskey & Bailey, 2005; O'Connor, 2007).

- Do not leave students out of the grading process.

Involve students; they can – and should – play key roles in assessment and grading that promote achievement (Guskey & Bailey, 2005; O'Connor, 2007).

According to Winger (2005), there are other criticisms of current grading practices which help provide a case for standards-based grading. First, the idea that grades interfere with learning because they provide the leverage to entice students to cooperate, but discourage students from taking chances. Second, grades measure what we value most (Winger, 2005). They measure a student's willingness to cooperate and work hard rather than an understanding of the content. He also believes that third, grades do not provide accurate feedback. When grades are not deliberately connected to learning, they provide little valuable feedback regarding a student's academic strengths and weaknesses and can be counter productive (Winger). If teachers expect grades to promote learning, then they must be sure that the grades assess and report the learning that they believe is most essential. Grade components must align with the state and district standards (Winger).

More recent solutions to problems with current grading practices are those that involve standards-based grading.

The promise of standards-based grading is that both teachers and students will have a clearer conception of what needs to be learned and what constitutes successful performance. This results in greater specification of what student-generated evidence is needed for evaluating the standard, how grades should be aligned to the evidence, and how effort and other "non academic" factors are reported. This should lead to less reliance on teacher impressions of student effort and improve the validity of grading. (Guskey, 2009, p. 107)

#### *Mandated Testing*

In the current climate of mandated testing, it is difficult to have a "civil discussion" about NCLB as proponents and dissenters weigh in. Douglas Reeves, a centrist on testing issues who heads the Center for Performance Assessment based in Denver, discusses the myths associated with this legislation. Reeves (2004) argues against the premise that this law is a Republican Party tactic to support vouchers and charter schools. His evidence is the Executive Order, signed by then President Bill Clinton, allowing parents to move their children out of schools failing to achieve adequate progress.

It is impossible to find someone who does not have an opinion about the current state of testing in public education. Despite the controversy, proponents of testing argue its merits. Reality Check (2002), a public opinion survey, reported that there is an across the board agreement that schools are moving forward with consideration to standards and testing, and as of yet no backlash has been initiated against the more rigorous requirements (Public Agenda, 2002). Dillon (2007) quoted Robert Linn, an education professor emeritus at the University of Colorado at Boulder and a frequent critic of NCLB who had reviewed results of the legislation and his comment stated, "I was a little surprised that things were generally as positive as they were, so it may be that I would say that NCLB is contributing more positively than I had given it credit for" (p. 7). His comments centered on a study of NCLB that he took part in by the Center on Education Policy (Dillion).

The language surrounding the aura of testing has been changing. In order to eliminate, as much as possible, the subjective nature in the determination of student achievement, state and district policymakers are making every effort to report performance in terms that are clear and understandable to students, parents, and the public

(Stapleman, 2000). As a result, students, parents and faculty are internalizing the "lingo" previously left to only the psychometricians to translate. It is now possible for the layperson to know and interpret individual achievement levels (Stapleman).

Those who support state-mandated standardized tests value these tests as tools in providing data and results necessary for schools to reform. Testing allows educators to focus instructional practices and to identify and abandon weak curriculum with the hope that eventually public education will turn to alternative forms of assessment (Schmoker, 2000). State tests are also powerful motivators for reform. Schools now have to set goals and evaluate their systems (Herman & Baker, 2005). The positive result to testing is its ability to focus on sub-groups and identify individual, particular needs because mandates also require these populations to meet AYP.

If nothing else NCLB has launched an unprecedented focus on the reading and math abilities of previously marginalized students. By requiring the desegregation of test scores by subgroups of students - such as English language learners, racial minorities, and students with special needs - NCLB ensures that



schools don't bury these students' test scores in schoolwide and gradewide averages or gloss over the achievement gaps that those scores reveal.

(Guilfoyle, 2006 p. 11)

For every proponent of standardized testing there is an equally vocal dissenter. Kohn is among the loudest critics. He stated,

Don't let anyone tell you that standardized tests are not accurate measures. The truth of the matter is they offer a remarkably precise method for gauging the size of the houses near the school where the test was administered. (Kohn, 2001, ¶ 1)

Kohn also argued there have been no positive effects of testing. He believed these tests are forcing good teachers out of education and forcing minority and low-income students out of school. Creativity is being stifled while "teaching is being narrowed and dumbed down, standardized and scripted" (Kohn, 2004, Dangers section ¶ 1). Other less emotional dissenters argue that test limitations, such as the multiple-choice format, does not indicate a student's ability to analyze in writing or apply processes (Shmoker, 2000). Arkansas has tried to overcome these limitations by providing questions which required

written responses and mathematical open response questions allowing students opportunities for application and inference (ADE, 2004).

Another critic, Popham (2007), compared using achievement tests to judge the quality of education to that of measuring temperature with a spoon: whereby achievement tests should only be used to make comparative interpretations. There is a fear that those who fund and evaluate schools will presume that poor scores indicate an inferior quality of education. It is this fear that may drive schools to lose creativity and spend time teaching the techniques of test taking rather than developing a more rigorous curriculum (Wallace, 2000). Furthermore, when the link between what is taught in the classroom and what is tested is ignored, negative results are likely to happen. In Texas, principals face the possibility of losing their jobs if their schools' standardized test scores don't measure up; superintendents can be fired and school boards can be dissolved if districts perform poorly (Bushweller, 1997).

There have been opponents of NCLB who see the school choice legislation as being one step closer toward a voucher system (Kohn, 2004). The most stringent critics believe the implication is the higher the student

achievement level, the more difficult the test becomes in order to ensure schools and students fail. As a result, public education will deteriorate, and school choice will allow the fulfillment of a conservative ideology whereby private education rules the day (Kohn, 2004).

NCLB presumed monitoring the percentage of students who are proficient in reading and mathematics this will be sufficient to identify schools that are doing a good job versus schools needing improvement (Cawelti, 2006). Unfortunately, this assumption has several flaws. First, because schools are held accountable for performance by student subgroups, large diverse schools are less likely to meet targets simply because they have more subgroups and hence more opportunities to miss achieving AYP goals (Nowak & Fuller, 2003).

Second, simply monitoring the percentage of students in a school who score at or above the proficient level in comparison with an annual target percentage places too much emphasis on student enrollment characteristics and any school that routinely receives a large influx of limited English proficient students each year will be at a disadvantage in comparison with a school that receives very few (Zehr, 2008). Third, monitoring school performances based on a single year assumes that current student

performance is a function of only the current year's instruction, ignoring past years. Fourth, reducing scores to a single cut-point, proficient or above versus below proficient, loses a significant amount of information about student performance (Thum, 2003). In most cases, a school will not receive credit for moving students up within an achievement level, nor will it be sanctioned if students move down within a level (Goldschmidt & Choi, 2007).

There is also the issue of test reliability as a test is gauged by the standard error of measurement or the degree to which the scores would spread out around the average score if the same student took the test many times (Crone, 2004). The measurement error on standardized tests can stem from a number of random factors, such as the student's health on the day of the test, the form of the test the student receives, or how well the student slept the night before. A mark of a well-designed test is that the measurement error is small relative to the range of scores on the test (Crone).

Another concern is that of test validity. Measurement experts are explicit about what makes a test valid in an accountability system. If alignment to the curriculum is weak and instruction does not match the standards, then the assessment will not meet the standards for validity and the

reported scores cannot be relied on as an adequate judge of a school's effectiveness (Popham, 2008). However, this is unfortunate when these scores are the determining factor in whether a school is rewarded or sanctioned (Barton, 2006). Popham (2008) argued that tests are not valid but referred to assessment validity which is defined as the accuracy of a score-based inference about a test taker's status. He stated, "Tests aren't valid or invalid; inferences are" (p. 82).

Early success reported by NCLB proponents may be an illusion if states are using statistical loopholes. If confidence intervals are used to calculate AYP where an error range is determined of a plus or minus it will skew the results (Popham, 2005). This statistical measure is correctly applied to sampling of a population and not on the complete population and providing an error range for an entire population who has already taken the test is statistically inappropriate. However, the federal government allows states to use this measure as way to keep lower numbers of schools in the needing improvement phase (Popham, 2005).

There are also less complex methods of using loopholes to fake AYP. Often cut scores seem arbitrary when states change them after raw scores have been reported or

weakening the rigor of a test by making items easier (Guifoyle, 2006). Furthermore, schools will often tutor bubble-students, those who fall just below the proficiency level, by using test taking techniques to move them upward. However, this does nothing to increase student achievement. In some cases, low-performing students are discouraged from attending on test day (Guilfoyle).

If accountability systems have the power to change behavior, as the early evidence indicates, then it is imperative to ensure that these systems change behavior in correct ways (Stecher, & Hamilton, 2002). However, sometimes high-stakes tests produce undesirable and unintended consequences, such as teaching to the test or excluding some students from testing (Fuhrman, 1999).

Positive consequences of mandated-testing for students may include better information about their own knowledge and skills. They may motivate students to work harder in school, send clearer signals to students about what to study, and help students associate personal effort with rewards (Reeves, 2006). Negative consequences for students might include tests frustrating and discouraging students from trying. They could potentially make students more competitive and cause them to devalue grades and school assessments. Tying assessments to students' graduation or

promotion can prompt students to drop out or increase the number of years necessary to graduate (Cawalti, 2006).

Positive consequences for teachers may include a more efficient way to diagnose individual student needs and help to identify areas of strength and weakness in the curriculum (Carter, 2006; Stiggins, 2008). Furthermore, testing could help identify content not mastered by students and help to redirect instruction (Carter; Stiggins). This will motivate teachers to work harder and smarter, lead teachers to align instruction with standards, and encourage teachers to participate in professional development to improve instruction (Carter; Stiggins).

Negative consequences for teachers may include possibly encouraging teachers to focus on specific test content more than curriculum standards (Zellmer, Frontier, & Pheifer, 2006). In a study of 376 elementary and secondary teachers in New Jersey, teachers indicated that they tended to teach to the test, often neglected individual students' needs because of the stringent focus on high stakes testing, had little time to teach creatively, and bored themselves and their students with practice problems as they prepared for standardized testing (Cawalti, 2006). This may lead teachers to engage in inappropriate test preparation, devalue teachers' sense of

professional worth, and entice teachers to cheat when preparing or administering tests (Zellmer, Frontier, & Pheifer).

Positive consequences for administrators may include an examination of school policies related to curriculum and instruction. Testing will help administrators to judge the quality of their programs, lead them to change school policies to improve curricula or instruction, and help them to make better resource allocation decisions (Carter, 2006; Reeves, 2006; Stiggins, 2008).

Negative consequences are leading administrators to enact policies to increase test scores but not necessarily increase learning. This may cause administrators to reallocate resources to tested subjects at the expense of other subjects and lead administrators to waste resources on test preparation (Stecher, & Hamilton, 2002).

Accountability models may also have unintended consequences. Schools in general must be careful to overcome a hazardous application of concentrating on the bubble kids. This is a practice which happens all too frequently and has become a negative, unintended consequence of testing.

This type of system may lead schools to employ selective discipline in an apparent attempt to shape



the testing pool, or even to utilize the school meals program to artificially boost student test performance by "carbo-loading" students for peak short-term brain activity. (Figlio, 2008 p. 25)

### *Summary*

The review of the literature indicated researchers are recognizing the need to link grading practices with standards and, subsequently the standardized assessments mandated by No Child Left Behind. How instructional practices and grading techniques influence student achievement, while not a new idea, was certainly worthy of continued study. Furthermore, there are a number of assessment types and it is essential to implement one linked to the goals set by the state. The assessment must be tied directly to state frameworks. It is imperative that a school understand that the district grading policies they currently have in place are effective, and if they are not resources must be used to implement change.

In chapter three the design methodology used to study middle school grading practices and the link to the Arkansas Benchmark Test was illustrated. Data was presented in chapter four which either proved or disproved the null hypothesis stated in chapter three. An analysis of the data

and its impending implications for assessment was also discussed in chapter five.

## CHAPTER THREE - DESIGN AND METHODOLOGY

### *Introduction*

With available research and data on instructional practices and the effect these strategies played on student achievement, it has been critical that a district understand whether its grading practices have been compatible with teaching techniques. The questions should be asked: Did the school teach the standards? Did the instructional practice best fit what was needed to increase student achievement? Did the state hold the school accountable by implementing an assessment which reflected the standards? And lastly, did the school have a grading policy which mirrored those previous points?

### *Subjects*

The secondary data information used in the study originated in a Northwest Arkansas Middle School. It was normally accessible to the researcher. Information was gathered from the sixth, seventh, and eighth grade student populations using semester grade cards and state-mandated testing data. This information was previously generated over a three year period in both math and literacy. All student information was kept anonymous for the purpose of

the research. Table 1 provided a breakdown of student demographics of the three years included in the study.

---

Table 1.

*Demographics*

---

Year	<u>2005-2006</u>	<u>2006-2007</u>	<u>2007-2008</u>
Total Student Enrollment	411	394	416
Percent Free/Reduced	49	58	50
Percent Special Ed.	08	10	11
Percent English Second Language	08	13	13
Percent White	70	75	74
Percent Hispanic	21	24	20

---

*Note:* From the National Office for Research, Measurement, and Evaluation Systems (2009).

---

*Sampling Procedure*

In order to provide consistency in the study of grades and testing, a random sample was not appropriate as interference from multiple schools with different grading practices, teacher quality, and curriculums would hinder the results. Furthermore, to investigate how Arkansas educators view the grading practices in their districts and or classroom, a survey was sent out across the state via e-mail.

For the correlation between the math and literacy grades and the spring 2006 Benchmark, all samples that were included participated in both semesters and the spring 2006 Benchmark test. The sample size for sixth grade in both literacy and math included 114 students. The seventh grade sample size was 122 for both math and literacy. The eighth grade sample size was 127 for literacy and 125 for math.

For the correlation between the math and literacy grades and the spring 2007 Benchmark, all samples that were included participated in both semesters and the spring 2007 Benchmark test. The sample size for the sixth grade included 108 for literacy and 115 for math. The seventh samples included 88 students for literacy and 101 for math. The eighth grade sample sizes were 106 for literacy and 122 for math

For the correlation between the math and literacy grades and the spring 2008 Benchmark, all samples that were included participated in both semesters and the spring 2008 Benchmark test. The sample size for the sixth grade included 138 students for literacy and 124 for math. The seventh grade included 122 for literacy and 119 for math. The eighth grade sample sizes were 110 for literacy and 113 for math.

#### *Research Setting*

For the purpose of this study, all math and literacy grades of the sample groups were examined. These grades were taken directly from grade cards, and both semesters were given equal weight. The Arkansas Benchmark is a standardized test and the setting has a more controlled environment where standardized procedures were followed to the letter of the law. All tests were administered in an appropriate setting with certified staff and specific time constraints.

#### *Research Design*

The study was designed to limit the sample size to include only students who participated both semesters and took the Benchmark test for each of the three years. The purpose was to limit the degree to which outside extraneous variables could influence the results. A correlational

analysis was applied because "it allows us to examine the degree to which two variables are interrelated (Runyon, Coleman, & Pittenger, 2000).

### *Questions*

There were three questions addressed in this study to conclusively answer the hypothesis.

1. What relationship exists between the first and second semester grades in three consecutive school years including 2006, 2007, and 2008 in the sixth, seventh, and eighth grades?
2. What relationship exists between semester grades and the spring benchmark examination in three consecutive school years including 2006, 2007, and 2008 in the sixth, seventh, and eighth grades?
3. What were area educators' attitudes concerning grading practices and the relationship to student achievement on the Arkansas Benchmark Test?

### *Independent Variable*

The independent variable in the study was the averaged semester grades in math and literacy for the sixth, seventh, and eighth grades during the 2005-2006, 2006-2007, and 2007-2008 school years.

*Dependent Variable*

The dependent variable in the study was the raw score percents on the Arkansas Benchmark Test in math and literacy for the sixth, seventh, and eighth grades during the 2005-2006, 2006-2007, and 2007-2008 school years.

*Hypotheses**Null Hypothesis*

The semester grades in literacy and math in a Northwest Arkansas Middle School was not an accurate predictor of student achievement on the Arkansas Benchmark Test.

*Alternate Hypothesis*

The semester grades in literacy and math in a Northwest Arkansas Middle School was an accurate predictor of student achievement on the Arkansas Benchmark Test.

*Procedure*

Grades over a three year period were examined starting with the 2005-2006 school year and ending with the 2007-2008 school year. The semesters for each of the sample populations were averaged. The Arkansas Benchmark Test was scored by the test manufacturer. The multiple choice questions were answered on a bubble sheet and ran through a scanning machine. The open responses were scored by trained individuals who used a rubric provided by the test



manufacturer and all responses were scored blind by multiple scorers. Results were returned to individual districts by May 31<sup>st</sup> of each year (ADE, 2004).

The first procedure planned in the study was to separately test the reliability of grading practices of the Northwest Arkansas Middle School used in the study. This allowed the researcher to determine if there would be a consistent outcome from semester to semester for the independent variable. A Pearson  $r$  correlation coefficient was calculated for each of the sixth, seventh, and eighth grades in both math and literacy for all three years.

For the primary measurement in the study designed to reject or accept the null hypothesis, a correlation coefficient was calculated between the averages of the semester grades to corresponding spring Benchmark assessment scores. Students' scale scores on the benchmark had been converted to raw score percents. This was repeated for the STAR Math pre-test in 2006 and the spring 2007 Benchmark assessment as well as the STAR Math pre-test in the fall of 2007 and the spring 2008 Benchmark assessment. Grades and test scores included each of the sixth, seventh, and eighth grade levels for both math and literacy.

The coefficient of determination was also figured to show the effect the independent variable, student grades,

had on the dependent variable, the spring Benchmark Assessment. Also, surveys were collected from around the state and the results were compiled to gather further information from Arkansas educators. Questions were designed to evaluate educators' views of their district's or school's grading practices.

### *Statistical Treatment of Data*

#### *Pearson $r$ Correlation*

The primary statistical measurement was the Pearson  $r$  correlation to study the relationship the independent variable had on the dependent variable. The independent variable in this study was the averaged semester grades for both math and literacy in grades six, seven and eight. The dependent variable was the student results on the Arkansas criterion-reference Benchmark Test. The correlation is one of the most common and most useful statistics. It is a single number that describes the degree of relationship between two variables (Trochim, 2008.) The correlation coefficient will vary in size from 0 to 1.00. A 0 indicates absolutely no relationship between the variables; a 1.00 indicates the strongest possible relationship. It is designated by the symbol  $r$

### *Coefficient of Determination*

Another technique used to interpret the correlation coefficient is to calculate the coefficient of determination. "The coefficient of determination tells us the percentage of variance in one variable that can be described or explained by the other variable" (Runyon, Coleman, & Pittenger, 2000). It is designated by the symbol  $r^2$ .

### *Summary*

Three years worth of data was accumulated and a correlation coefficient was calculated by using Pearson  $r$ . A quasi-experimental design allowed for multiple variables and multiple measures. Furthermore, reliability tests were performed on the semester grades in both math and literacy. In addition, nuisance variables were considered and limited to the best of the researcher's ability.

All data was run through the SPSS Graduate Pack software to reduce potential calculation errors. As a note, when calculating correlations it was necessary to distinguish this will not provide the researcher a causal relationship, but instead only measures them to look for relations between a set of variables. In chapter four the data was analyzed and in chapter five the results were

examined and provided implications and recommendations for schools.

## CHAPTER FOUR - RESULTS

### *Introduction*

The data was compiled and the primary test, a Pearson  $r$  correlation, was performed on three grade levels for three separate years in both literacy and math. The average of first and second semester grades was correlated with the corresponding average raw score percent.

### *Results*

The null hypothesis for this study stated that semester grades in literacy and math in a Northwest Arkansas Middle School was not an accurate predictor of student achievement on the Arkansas Benchmark Test. The results were mixed over the three-year period. After examining both the correlation coefficients and the coefficients of determination, the null hypothesis was accepted.

### *Analysis of Data*

*Research question number one.* What relationship exists between the first and second semester grades in three consecutive school years including 2006, 2007, and 2008 in the sixth, seventh, and eighth grades?

The first step in the study was to examine whether there was any correlation between the first and second semester grades. This was necessary to ensure the average of the two semesters would be viable as an independent variable for the study as a whole. The correlations in Table 2 and Table 3 showed that over the three years in the three grades for both math and literacy were statistically significant. The range in literacy was .709 to .877 with an average of .807. The range in math was .765 to .858 with an average of .819.

---

Table 2.

*Correlation for 1st and 2nd Semester Literacy Grades*

---

Grade	2005-2006	2006-2007	2007-2008
	Coefficients	Coefficients	Coefficients
Sixth	.877 n=114	.871 n=108	.804 n=138
Seventh	.763 n=122	.831 n=88	.736 n=122
Eighth	.823 n=127	.848 n=106	.709 n=110

---

Note: n=student sample size

Correlation significant at the .500 level

---

Table 3.

*Correlation for 1st and 2nd Semester Math Grades*


---

Grade	2005-2006	2006-2007	2007-2008
	Coefficients	Coefficients	Coefficients
Sixth	.765 n=114	.844 n=115	.820 n=124
Seventh	.835 n=122	.824 n=101	.858 n=119
Eighth	.790 n=125	.822 n=122	.811 n=124

---

Note: n=student sample size

Correlation significant at the .500 level

---

*Research question number two.* What relationship exists between semester grades and the spring benchmark examination in three consecutive school years including 2006, 2007, and 2008 in the sixth, seventh, and eighth grades?

*Pearson r*

The results were mixed over the three years and three grade levels in literacy. The coefficients ranged from .296

to .788 with an average of .454. Only two of the nine calculations were over the .500 target for acceptance.

The results were also mixed in math over the three years and three grade levels. However, the range was not as wide-spread; the coefficients ranged from .448 to .741 with an average of .553. Five of the nine calculations were over the .500 target for acceptance. Tables 4 and 5 displayed the results of the calculations.

---

Table 4.

*Correlation of Averaged Semester Grades and Benchmark Raw Score Percents in Literacy*

---

Grade	2005-2006	2006-2007	2007-2008
	Coefficients	Coefficients	Coefficients
Sixth	.338 n=114	.788 n=108	.508 n=138
Seventh	.364 n=122	.584 n=88	.296 n=122
Eighth	.422 n=127	.452 n=106	.335 n=110

---

Note: n=student sample size

Correlation significant at the .500 level

---



Table 5.

*Correlation of Averaged Semester Grades and Benchmark Raw Score Percents in Math*

---



---

Grade	2005-2006	2006-2007	2007-2008
	Coefficients	Coefficients	Coefficients
Sixth	.448 n=114	.635 n=115	.433 n=124
Seventh	.583 n=122	.648 n=101	.741 n=119
Eighth	.567 n=125	.452 n=122	.475 n=113

---

Note: n=student sample size

Correlation significant at the .500 level

---

*Coefficient of Determination*

The coefficient of determination's range in literacy was wide-spread as were the Pearson *r* calculations. The actual coefficients were converted from raw numbers to percents and the scope was from 11% to 62% with an average of 22.67%. The coefficient of determination provided for the strength of the correlation, and, as a result, the correlation between the averaged semester grades and the average raw score percents of the benchmark were weak.

The coefficient of determination's range in math was not as wide-spread as was for literacy. The scope was from 19% to 55% with an average of 31.67%. The coefficient of determination provided for the strength of the correlation; as a result, the correlation between the averaged semester grades and the average raw score percents of the benchmark were weak. The results of these calculations were displayed in Tables 6 and 7.

---

Table 6.

*Coefficient of Determination of Averaged Semester Grades and Benchmark Raw Score Percents in Literacy*

---

Grade	2005-2006	2006-2007	2007-2008
	Coefficients	Coefficients	Coefficients
Sixth	11%	62%	26%
Seventh	13%	34%	09%
Eighth	18%	20%	11%

---

*Note:* Coefficient of Determination is represented in percents

---

Table 7.

*Coefficient of Determination of Averaged Semester Grades  
and Benchmark Raw Score Percents in Math*

---

Grade	2005-2006	2006-2007	2007-2008
	Coefficients	Coefficients	Coefficients
Sixth	20%	40%	19%
Seventh	34%	42%	55%
Eighth	32%	20%	23%

---

*Note:* Coefficient of Determination is represented in percents

---

*Research question number three.* What were area educators' attitudes concerning their grading practices and their relationship to student achievement on the Arkansas Benchmark Test?

*Survey*

The survey results were made up of ninety-six responses with forty-five different schools represented. Overall, educators responded to the high-end of the scale when answering questions about their particular grading

practices and their grasp of assessment types. However, they marked at the lower-end of the scale when answering questions about the use of rubrics or scoring guides and questions concerning professional development or training connected with grading practices. The majority of respondents fell in the mid-range when answering the specific question about their grades and the connection they may or may not have to the benchmark. The results were displayed in chart form and by building in Appendix A. The questions and responses were as follows;

- Question # 1: I purposely consider the effect an individual grade will have on the overall nine-week or semester grade of a student. The majority of responses were in the almost always category.
- Question # 2: I understand how a grade may influence the overall grade. The majority of responses were in the almost always category. The majority of responses were in the almost always category.
- Question # 3: I pre-determine the number of total points I will have in a nine-weeks. The majority of responses were in the rarely category.

- Question # 4: I believe there has been sufficient professional development concerning grading. The majority of responses were in the rarely category.
- Question # 5: I feel I have had sufficient training on how to develop and grade using rubrics or scoring guides. The majority of responses were in the rarely category.
- Question # 6: I use a scoring guide or rubric on more than 75% of my assignments. The majority of responses were in the frequently category.
- Question # 7: I consider the assessment and the standards before I plan the activities. The majority of responses were in the almost always category.
- Question # 8: I give extra-credit points during a nine-weeks grading period. The majority of responses were in the rarely category.
- Question # 9: If so, is the total percent more than 10% of the overall grade. The majority of responses were in the never category.
- Question # 10: I understand the difference between grades and assessment. The majority of responses were in the always category.

- Question # 11: I use the following types of assessment in my classroom
  - A. Assessment for learning: The majority of responses were in the frequently category.
  - B. Assessment of learning. The majority of responses were in the almost always category.
  - C. Formative assessment. The majority of responses were in the frequently category.
  - D. Summative assessment. The majority of responses were in the frequently category.
  - E. Interim assessments. The majority of responses were in the frequently category.
  - F. Pre-assessment. The majority of responses were in the frequently category.
- Question # 12: I feel the grade a student earns in my class is indicative of the rating a student will earn on the benchmark. The majority of responses were in the frequently category.
- Question # 13: I curve grades in my class. The majority of responses were in the rarely category.
- Question # 14: I give credit/no credit grades in my class. The majority of responses were in the never category.

- Question # 15: I know what an A grade response should look like on any assignment. The majority of responses were in the almost always category.

Question number twelve was the imbedded question to see exactly how respondents believed their grading practices correlated to the benchmark. Of ninety-seven educators surveyed, eighty-two percent fell in the always to frequently categories. Only eighteen percent fell in the rarely to never categories. Respondents had a more positive belief that grades reflected benchmark scores than the Pearson  $r$  correlation coefficients indicated.

#### *Deductive Conclusions*

Based on the mixed results, it was impossible to determine a statistically significant correlation between students' averaged semester grades and the corresponding raw score percent on the spring administration of the Arkansas Benchmark Test. The original null hypothesis stated that the averaged semester grades were not an accurate predictor of student achievement on the Arkansas Benchmark Test, and after the analysis was completed, the null hypothesis was accepted.

*Summary*

After a Pearson  $r$  and a coefficient of determination were calculated, the results of the study indicated a weak relationship between the averaged semester grades and the corresponding raw score percents on the Arkansas Benchmark test. The correlations were too wide-spread to discover any concrete patterns. However, the correlations were higher when math semester grades were compared to math Benchmark scores. In chapter five the implications and recommendations for effective schools were discussed.



## CHAPTER FIVE - DISCUSSION

### *Introduction*

The rationale for the study was to help Arkansas districts achieve Adequate Yearly Progress as state funding was directly linked to test scores. Districts accurately predicting student achievement can target student weakness prior to benchmark testing and focus efforts on direct remediation. If target areas were identifiable, districts could restructure their curriculum more effectively and efficiently.

There has been a virtual revolution in assessment practices due in part to NCLB; however, the evolution of grading practices was much slower. Sometimes grading systems currently in place subvert the good intentions of reformed assessment systems. It was imperative that grades reflect the achievement levels arrived at in these criterion-referenced tests required under the law. Schools have been concentrating on grading systems to accurately reflect student progress and how accurately they meet state standards throughout the year. This was important because

in many states reporting of annual scores are delivered too late in the year to accurately remediate student weaknesses. While the state tests have been important communicators of student achievement and they allow schools to reform curriculum and instruction long-term, they did not impart ongoing information that schools needed to incrementally improve instructional programs.

As discussed in chapter four, the first step was to test the reliability of the independent variable used in the study. Once it was determined that the independent variable produced a consistent outcome over time the correlation of the variables were calculated. The degree to which the independent variable, the averaged semester grades, had on the dependent variable, the Arkansas Benchmark Test, was also measured. This was done by calculating the Pearson  $r$  correlations. The Coefficient of Determination was also factored to determine the strength of the correlation coefficients.

The samples were not chosen by random in order to limit the nuisance variables. Only students who had completed both semesters and the current years benchmark were considered. This eliminated as much as possible outside curriculums and instructional practices.

Relationships to sub-populations were not given any more or less consideration to the sample population

#### *Implication for Effective Schools*

The results of the study showed that further examinations of school grading practices are necessary. Schools must train teachers on effective grading and its relationship to student achievement. There must be a link between instructional practices, assessment, and the ensuing grades. Practices that do not serve the students' best interests are not necessary and should be abolished.

#### *Recommendations*

The survey of area educators showed that they believed professional development on grading was needed and the results of the research supported that conclusion. Embedded professional development is the most effective way to implement changes within a school system. The researcher recommends schools would benefit from the implementation of a professional learning community. Dufour offers several books that are available to help a district establish a professional learning community. Stiggins also provides research-based information to help teachers better use formative assessment techniques in the classroom. Ultimately, the goal must be to eliminate grading practices that do not accurately reflect student achievement.

*Summary*

Studies such as this provide a district with valuable information. In the age of accountability, a district must know if the practices in place tie directly to student achievement. If they do not, they must be eliminated and replaced with research-based practices that do enhance or emulate growth for individual students.

## REFERENCES

- Airasian, P.W. (2000). *Assessment in the classroom: A concise approach* (2nd ed.). New York, NY: McGraw-Hill. 195-200.
- Arkansas Department of Education. (2004, June). *Rules governing the Arkansas comprehensive testing, assessment and accountability program*. ADE 188-1, Retrieved January 23, 2009 from <http://arkansased.org>
- Arkansas Department of Education. (2005, July). *Consolidated state application accountability plan*. ADE News release, Retrieved January 23, 2009 from <http://arkansased.org>
- Arkansas Department of Education. (2007, September). *Arkansas students performance holds steady on the "Nation's Report Card"*. [Electronic version]. ADE News release, Retrieved January 23, 2009 from <http://arkansased.org>
- Arkansas Department of Education. (2008a). *ACTAAP: Arkansas alternate portfolio assessment for students with disabilities for grades 3-8,11*. ADE Handbook, Retrieved October 25, 2008 from <http://arkansased.org>

Arkansas Department of Education, (2008b, February).

*Consolidated state application accountability plan.*

Workbook, Retrieved January 23, 2009 from <http://>

[arkansased.org](http://arkansased.org)

Association of American Publishers. (2000). *Standardized*

*assessment: A primer* (revised edition.) [Brochure].

Washington, DC.

Baker, E., Linn, R., & Herman, J. (2002, Winter). *Standards*

*for educational accountability systems.* (Policy Brief

5). Los Angeles: University of California, National

Center for Research on Evaluation, Standards, and

Student Testing.

Barnwell, P. (2008, June). Could standard grading practices

be counterproductive? *Education Week*, 27(41), 1, 22.

Barton, P. (2006, November). Needed: Higher standards for

accountability. *Education Leadership*, 64(3), 28-31.

Bond, L. A. (1996). Norm- and criterion-referenced testing.

*Practical Assessment, Research and Evaluation*, ISSN

1531-7714. Retrieved August 21, 2006 from

<http://pareonline.net>

- Bushweller, K. (1997). Teaching to the test: Increasingly, schools are finding it just makes sense to align curriculum and assessment. *American School Board Journal*, Retrieved August 29, 2006 from <http://www.asbj.com/achievement/aa/aa4>
- Carter, L. (2007). *Total instructional alignment: From standards to student success*. Bloomington, IN: Solution Tree.
- Cavanaugh, S. (2007, June). State tests, NAEP often a mismatch. *Education Week*, 26(41), 1,23.
- Cavanaugh, S. (2008, June). Since NCLB law: Test scores on the rise. *Education Week*, 27(43), 1,16.
- Cawelti, G. (2006, November). The side effects of NCLB. *Education Leadership*, 64(3), 64-68.
- Chappuis, S., & Chappuis, J. (2008, January). The best value in formative assessments. *Education Leadership*, 5(4), 14-18.
- Christopher, S. (2008, January). Homework: A few practice arrows. *Educational Leadership*, 65(4), pp. 74-75.
- Cicchinelli, L., Gaddy, B., Lefkowitz, L., & Miller, K. (2003, April). *No child left behind: Realizing the vision*. (Policy Brief). Aurora, CO: Mid-continent Research for Education and Learning.

- Clarke, M., Madaus, G., Horn, C., Ramos, M., Lynch, C.A., & Lynch, P.S. (2001, April). The marketplace for educational testing. *Educational Testing*, 2(3), Retrieved August 28, 2006 from <http://www.bc.edu/research/nbetpp/publications>
- Clymer, J., & Wiliam, D. (2007, January). Improving the way we grade science. *Educational Leadership*. 64(4) 36-42.
- Crone, T. (2004, September). What test scores can and cannot tell us about the quality of our schools. *Business Review*, Philadelphia, PA: Federal Reserve Bank. Retrieved November 18, 2008 from <http://www.philadelphiafed.org/research-and-data/publications/business-review/2004/q3/brq304tc.pdf>.
- Davis, B. (1999) *Grading practices*. (Tools for Teaching). Berkley, CA: Office of Educational Development.
- Dillon, S. (2007, June). New study finds gains since no child left behind. *The New York Times*.
- Elmore, R. (2003, November). A plea for strong practice. *Education Leadership*, 61(3), 6-10.
- Figlio, D. (2008). Testing and accountability in the NCLB era. *Education Week*, Retrieved August 26, 2008 from [www.edweek.org](http://www.edweek.org)



- Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge no child left behind? *Educational Researcher*, 36(5) 268-278.
- Fuhrman, S. (1999). *The new accountability*, (CPRE Policy Briefs), Philadelphia PA: Consortium for Policy Research in Education.
- Gillum, J., Rivera, A., Sanchez, G., & Younger, J. (2008, May 12). Grade inflation adds to woes, especially in middle schools. *Arizona Daily Star*, Retrieved on October 3, 2008, from [www.dailystar.com](http://www.dailystar.com)
- Goldschmidt, P., & Choi, K. (2007, Winter). *The practical benefits of growth models for accountability and the limitations under NCLB*. (Policy Brief 9), Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Guilfoyle, C. (2006, November). NCLB: Is there life beyond testing? *Education Leadership*, 64(3), 8-13.
- Guskey, T.R., & Bailey, J.M. (2001). *Developing grading and reporting systems for student learning*. Thousand Oaks, CA: Corwin Press.
- Guskey, T. (2008, January). The rest of the story. *Education Leadership*, 65(4), 28-34.

- Guskey, T., (2009). *Practical solutions for serious Problems in standards-based grading*. Thousand Oaks, CA: Corwin Press.
- Harris, D., & Carr, J. (1996). *How to use standards in the classroom*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Herman, J.L., & Baker, E.L. (2005, November). Making benchmark testing work. *Educational Leadership*, 63(3), 48-54.
- Herman, J.L., Osmundson, E., Ayaly, C., Schneider, S., & Timms, M. (2006, December). *The nature and impact of teacher's formative assessment practices*. (CSE Technical Report 703). Los Angeles, Ca: National Center for Research on Evaluaton, Standards, and Studert Testing,
- Hoff, D. J. (2008, July). 2 new coalitions seek influence on campaigns. *Education Week*, 27(42), 1, 24.
- Huhn, C. (2005, November). How many points is this worth? *Educatonal Leadership*, 61(3), 36-40.
- Johnson, J. (2003, November). What does the public say about accountability? *Educatonal Leadership*, 63(3), 83-84.

- Junker, B., Weisburg, Y., Matsumura, L.C., Crosson, A., Wolf, M. K., Levison, A., & Resnick, L. (2006, January). *Overview of the instructional quality assessment*. (CSE Technical Report 671). Los Angeles, Ca: National Center for Research on Evaluation Standards, and Student Testing.
- Kennedy-Manzo, K. (2008, June). Principals' group calls for national academic standards and tests. *Education Week*, 27(41), 6.
- Kennedy-Manzo, K. (2008, August). Achieve finds common core of standards in states. *Education Week*, 27(45), 6.
- Klein, A. (2008, June). Kennedy's illness raises doubts for NCLB. *Education, Week*, 27(41), 17-18.
- Klien, S., Hamiltion, L., McCaffrey, D., Stecher, B., Robyn, A., & Burroughs, D. (2000) *Teaching practices and student achievement*. (Report of First-Year Findings from the 'Mosaic' Study of Systemic Initiatives in Mathematics and Science). Rand Corporation. Retrieved on September 23, 2008 from [http://www.rand.org/pubs/monograph\\_reports/MR1233/index.html](http://www.rand.org/pubs/monograph_reports/MR1233/index.html)
- Kohn, A. (1989). From degrading to de-grading. *The High School Magazine*, 6(5).

- Kohn, A. (2001, January). Beware of the standards, not just the tests. *Education Week*, Retrieved from <http://www.alfiekohn.org/teaching/edweek>
- Kohn, A. (2003, November). The dangerous myth of grade inflation. *The Chronicle of Higher Education*, Retrieved November 01, 2008 from <http://www.alfiekohn.org>
- Kohn, A. (2004, April). Test today, privatize tomorrow: Using accountability to "reform" public schools to death. *Phi Delta Kappan*, Retrieved April 1, 2005 from <http://www.alfiekohn.org/articles.htm>
- Laitsch, D. (2005, July). *A policymaker's primer on testing and assessment*. (Infobrief 42), Alexandria, VA: Association for Supervision and Curriculum Development.
- Lewis, A. (2000, April). *High-stakes testing: Trends and issues*. (Policy Brief), Aurora CO: Mid-Continent Research for Education and Learning.
- Linn, R. (1998, April). Assessments and accountability. *Educational Researcher* 29(2), 4-16.
- Loup, C., & Peterilli, M., (2005, Winter). *Crystal apple: Education insider's predictions for no child left behind's reauthorization*. (Research Brief), Washington, D.C.: Thomas Fordham Institute.

- Mandel, S. (2006, March). What new teachers really need. *Educational Leadership*, 63(6), 66-69.
- Marsh, J., Pane, J., & Hamilton, L. (2006). Making sense of data-driven decision making in education. *Occasional Paper*, Rand Education
- Marzano, R. (2000). *Transforming classroom grading*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Marzano, R. (2006). *Classroom assessment and grading that work*. Alexandria, VA: Association for Supervision and Curriculum Development.
- McTighe J. & O'Connor, K. (2005, November). Seven practices for effective learning. *Educational Leadership*, 63(3), 10-17.
- Nowak, J., & Fuller, B. (2003). *Penalizing diverse schools? Similar test scores, but different students bring federal sanction*. (Pace Policy Brief), Berkley CA: Policy Analysis for California Education.
- O'Connor, K. (2007). *A repair kit for grading: 15 fixes for broken grades*. (1st ed.). Portland, OR: Educational Testing Service.
- Olson, A. (2007, January). Growth measures for systemic change. *The School Administrator*, 64(1), 10-16.

- Pascopella, A. (2006, January). Nitty-gritty data. *The District Administrator*, 42(1), 36-41.
- Pollard, E. (2008, October). Multiple yardsticks for measuring proficiency. *Education Week*. Retrieved March 29, 2008 from <http://www.edweek.org//search.html?prx=p&occ=p&qsl=NAEP+&prd=y&srt=r&src=63&idx=15>
- Popham, J. W. (2003, December). Living (or dying) with your NCLB tests. *The School Administrator*, Retrieved July 5, 2006, from the American Association of School Administrators' database.
- Popham, J.W. (2006a, September). Content standards: The unindicted co-conspirator. *Educational Leadership*, 64(1), 87-88.
- Popham, J.W. (2006b, November). Phony formative assessments: Buyer beware. *Educational Leadership*, 64(3), 86-87.
- Popham, J.W. (2007). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Popham, J.W. (2008, September). A misunderstood grail. *Education Leadership*, 66(1), 82-83.
- Public Agenda Online. (2002). *Reality Check 2002*. Retrieved August 31, 2006, from <http://publicagenda.org/specials/rcheck2002/htm>

- Rabinowitz, S. (2001, December). Balancing state and local assessments: A district's duty to meet needs of all students through testing. *The School Administrator*, Retrieved July 5, 2006, from the American Association of School Administrators database.
- Ravitch, D. (2007). *Ed speak: A glossary of education terms, phrases, buzzwords, and jargon*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Reeves, D.B. (2004). *Accountability for learning: How teachers and school leaders can take charge*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Reeves, D.B. (2006, November). Preventing 1000 failures. *Educational Leadership* 64(3) 88-89.
- Reeves, D.B. (2008, February). Effective grading. *Educational Leadership* 65(5) 85-87.
- Runyon, R.P, Coleman, K.A., & Pittenger, D.J. (2000). *Fundamentals of behavioral statistics*. Boston, MA: McGraw Hill.
- Sanders, W., & Horn, S. (1995, March). Educational assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes.

*Education Policy Analysis Archives*, 3(6).

Sausner, R. (2005, August). Making assessments work.

*District Administrator*. 41(8), 31-34.

Schmoker, M. (2000, February). The results we want.

*Educational Leadership*, 57(5), 62-65.

Scriffiny, P. (2008, October). Seven reasons for standards-

based grading. *Educational Leadership*, 66(2), 70-74.

Sokola, D.P., Weinberg, H.M., Andrzejewski, R.J., & Doorey,

N.A. (2008, May). Fixing the flaw in the growth

model. *Education Week*, 27(38), 26-27, 29.

Stapleman, J. (2000). *Standards-based accountability*

*systems*. (Policy Brief), Aurora CO: Mid-Continent

Research for Education and Learning.

Starkman, N. (2006, September). Building a better student.

*The Journal*, 33(14), 41-46.

Stecher, B., & Hamilton, L. (2002). Putting theory to the

test: Systems of "educational accountability" should

be held accountable. Rand Corporation. Retrieved

September 26, 007 from <http://www.rand.org/>

[publications/randreview](http://www.rand.org/publications/randreview)

Stiggins, R. (2008, April). *A call for the development of*

*balanced assessment systems*. (Assessment Manifesto).

Portland, OR: ETS Assessment Training Institute.



- Thum, Y. M. (2003). *No Child Left Behind: Methodological challenges and recommendations for measuring adequate yearly progress*. (CSE Technical Report 590). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards, and Student Testing.
- Toch, T. (2006, November). Turmoil in the testing industry. *Education Leadership*, 64(3), 53-56.
- Tomlinson, C. (2008, January). Leaning to love assessment. *Education Leadership*, 65(4), 8-13.
- Trochim, W. M. (2008). *The research knowledge base*. (2<sup>nd</sup> Edition). Retrieved January 09, 2009 <http://www.socialresearchmethods.net/kb/>
- Trumbell, E., & Farr, B. (2000). *Grading and reporting student progress in an age of standards*. Norwood: Christopher-Gordon Publishers, Inc.
- United States Department of Education (DOE). (2001, January). *No Child Left Behind (NCLB)*. Public Law 107-110 Retrieved on February 24, 2009, from <http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf>
- United States Department of Education. (2003, February). *Standards and assessments*. Title I Director's Conference, Retrieved September 5, 2006, from

<http://www.ed.gov>

- Wallace, D. (2000, February). Results, results, results? *Educational Leadership*, 57(5), 66-68.
- Wiggins, G., & McTighe, J. (2005) *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Wilcox, J. (2007, February). NCLB on the eve of reauthorization: Calls for fundamental overhaul greet new congress. *Education Update*, 49(2).
- Winger, T. (2005, November). Grading to communicate. *Educational Leadership*, 63(3), 61-65.
- Wong, K., & Nicotera, A. (2007). *Successful schools and educational accountability: Concepts and skills to meet leadership challenges*. Boston, MA: Pearson Education Inc.
- Wright, D., & Wise, M.J. (1988). Teacher judgement in student evaluation: A comparison of grading methods. *Journal of Educational Research*, 82(1), 10-14.
- Zehr, M. (2008, July). States struggle to meet achievement standards for ELLs. *Education, Week*, 27(43), 12.
- Zellmer, M., Frontier, A., & Pheifer, D. (2006, November). What are NCLB's instructional costs? *Educational Leadership*, 64(3), 43-46.

## Appendix A

Table A1

*Survey Results of Area Educators*

Question	Always	Almost Always	Frequently	Rarely	Never
1. I purposely consider the effect an individual grade will have on the overall nine-week or semester grade of a student	19	33	30	20	5
2. I understand how a grade may influence the overall grade.	61	31	19	1	0
3. I pre-determine the number of total points I will have in a nine-weeks.	2	8	15	43	31
4. I believe there has been sufficient professional development concerning grading.	4	12	9	60	15
5. I feel I have had sufficient training on how to develop and grade using rubrics or scoring guides.	16	28	23	32	2
6. I use a scoring guide or rubric on more than 75% of my assignments.	10	18	40	30	3
7. I consider the assessment and the standards before I plan the activities.	31	41	22	8	0
8. I give extra-credit points during a nine-weeks grading period.	1	10	23	40	25
9. If so, is the total percent more than 10% of the overall grade.	0	1	0	31	58
10. I understand the difference between grades and assessment.	50	36	10	3	1

11. I use the following types of assessment in my classroom					
A. Assessment for learning	19	35	37	4	2
B. Assessment of learning	24	36	34	0	1
C. Formative assessment	21	27	43	6	2
D. Summative assessment	25	28	41	4	1
E. Interim assessments	14	18	48	15	1
F. Pre-assessments	10	17	37	31	3
12. I feel the grade a student earns in my class is indicative of the rating a student will earn on the benchmark.	3	27	50	16	1
13. I curve grades in my class.	0	1	4	46	44
14. I give credit/no credit grades in my class.	5	1	18	34	37
15. I know what an A grade response should look like on any assignment.	38	45	14	2	0

Table A2.

*Survey Results - Elementary*

Question	Always	Almost Always	Frequently	Rarely	Never
1. I purposely consider the effect an individual grade will have on the overall nine-week or semester grade of a student.	10	4	11	8	3
2. I understand how a grade may influence the overall grade.	23	9	4	0	0
3. I pre-determine the number of total points I will have in a nine-weeks	2	2	5	13	12
4. I believe there has been sufficient professional development concerning grading.	1	0	5	23	6
5. I feel I have had sufficient training on how to develop and grade using rubrics or scoring guides.	7	6	10	14	0
6. I use a scoring guide or rubric on more than 75% of my assignments.	5	7	11	11	1
7. I consider the assessment and the standards before I plan the activities.	14	17	5	0	0
8. I give extra-credit points during a nine-weeks grading period.	0	4	6	13	10
9. If so, is the total percent more than 10% of the overall grade.	0	0	0	10	18
10. I understand the difference between grades and assessment.	24	9	3	0	0
11. I use the following types of assessment in my classroom	12	13	8	0	1
A. Assessment for learning					
B. Assessment of learning	11	13	9	0	0
C. Formative assessment	10	10	13	1	1
D. Summative assessment	11	4	18	1	1
E. Interim assessments	9	5	16	2	1
F. Pre-assessments	6	8	11	6	1
12. I feel the grade a student earns in my class is indicative of the rating a student will earn	3	10	20	1	0

on the benchmark.					
13. I curve grades in my class.	0	0	2	11	21
14. I give credit/no credit grades in my class.	2	0	5	8	17
15. I know what an A grade response should look like on any assignment.	18	15	1	0	0

Table A3

*Survey Results - Intermediate*

Question	Always	Almost Always	Frequently	Rarely	Never
1. I purposely consider the effect an individual grade will have on the overall nine-week or semester grade of a student.	2	4	6	1	0
2. I understand how a grade may influence the overall grade.	5	4	2	1	0
3. I pre-determine the number of total points I will have in a nine-weeks	0	1	0	6	6
4. I believe there has been sufficient professional development concerning grading.	1	5	1	5	1
5. I feel I have had sufficient training on how to develop and grade using rubrics or scoring guides.	3	5	3	2	0
6. I use a scoring guide or rubric on more than 75% of my assignments.	1	4	6	2	0
7. I consider the assessment and the standards before I plan the activities.	5	4	3	1	0
8. I give extra-credit points during a nine-weeks grading period.	0	0	4	7	2
9. If so, is the total percent more than 10% of the overall grade.	0	0	0	4	7
10. I understand the difference between grades and assessment.	7	4	0	1	0
11. I use the following types of assessment in my classroom	3	3	7	0	0
A. Assessment for learning					
B. Assessment of learning	4	3	6	0	0
C. Formative assessment	5	4	4	0	0
D. Summative assessment	4	5	3	0	0
E. Interim assessments	2	3	7	1	0
F. Pre-assessments	1	1	6	5	0
12. I feel the grade a student earns in my class is indicative of the rating a student will earn	0	4	8	1	0

on the benchmark.					
13. I curve grades in my class.	0	0	0	8	5
14. I give credit/no credit grades in my class.	0	1	1	6	4
15. I know what an A grade response should look like on any assignment.	6	5	1	1	0



Table A4

*Survey Results - Middle School*

Question	Always	Almost Always	Frequently	Rarely	Never
1. I purposely consider the effect an individual grade will have on the overall nine-week or semester grade of a student.	3	11	3	4	1
2. I understand how a grade may influence the overall grade.	11	7	4	0	0
3. I pre-determine the number of total points I will have in a nine-weeks	0	2	5	10	5
4. I believe there has been sufficient professional development concerning grading.	1	2	2	12	4
5. I feel I have had sufficient training on how to develop and grade using rubrics or scoring guides.	2	10	4	6	0
6. I use a scoring guide or rubric on more than 75% of my assignments.	3	3	10	6	0
7. I consider the assessment and the standards before I plan the activities.	6	8	4	3	0
8. I give extra-credit points during a nine-weeks grading period.	0	4	4	6	8
9. If so, is the total percent more than 10% of the overall grade.	0	0	0	7	12
10. I understand the difference between grades and assessment.	8	10	3	1	0
11. I use the following types of assessment in my classroom	3	7	8	2	0
A. Assessment for learning					
B. Assessment of learning	4	9	8	0	0
C. Formative assessment	3	7	9	3	0
D. Summative assessment	7	5	7	2	0
E. Interim assessments	2	6	6	7	0
F. Pre-assessments	3	5	8	5	1
12. I feel the grade a student earns in my class is indicative of the rating a student will earn	0	4	9	8	1

on the benchmark.					
13. I curve grades in my class.	0	0	0	8	13
14. I give credit/no credit grades in my class.	1	0	2	8	10
15. I know what an A grade response should look like on any assignment.	8	9	5	0	0

Table A5

*Survey Results - High School*

Question	Always	Almost Always	Frequently	Rarely	Never
1. I purposely consider the effect an individual grade will have on the overall nine-week or semester grade of a student.	4	9	10	7	1
2. I understand how a grade may influence the overall grade.	11	11	9	0	0
3. I pre-determine the number of total points I will have in a nine-weeks	0	3	5	15	8
4. I believe there has been sufficient professional development concerning grading.	1	5	1	20	4
5. I feel I have had sufficient training on how to develop and grade using rubrics or scoring guides.	4	8	6	10	2
6. I use a scoring guide or rubric on more than 75% of my assignments.	1	4	13	11	2
7. I consider the assessment and the standards before I plan the activities.	5	12	10	4	0
8. I give extra-credit points during a nine-weeks grading period.	1	2	9	14	5
9. If so, is the total percent more than 10% of the overall grade.	0	1	0	10	20
10. I understand the difference between grades and assessment.	11	13	5	1	1
11. I use the following types of assessment in my classroom	1	12	14	2	1
A. Assessment for learning					
B. Assessment of learning	6	11	11	0	0
C. Formative assessment	3	6	17	2	1
D. Summative assessment	3	14	13	1	0
E. Interim assessments	1	4	19	5	0
F. Pre-assessments	0	3	12	15	0
12. I feel the grade a student earns in my class is indicative of the rating a student will earn	0	9	13	6	0

on the benchmark.					
13. I curve grades in my class.	0	1	2	19	9
14. I give credit/no credit grades in my class.	2	0	10	12	6
15. I know what an A grade response should look like on any assignment.	6	16	7	1	0

## APPENDIX B

*E-mail to Area Educators*

I am asking you to respond to the following survey. All results will remain anonymous and the information will be tabulated as a whole to provide statistical data for my doctoral dissertation. The information you share is not designed for any other purpose.

Please leave the survey in the folder located in your workroom, or send it through inter-office mail. Please remember to check which building you are in, but do not put your name on the response sheet.

As always your help and effort is appreciated. Please call me at 423-4512 or email me at [msummers@bobcat.k12.ar.us](mailto:msummers@bobcat.k12.ar.us) if you have any questions.

Sincerely,

Matt Summers  
Middle School Principal  
Berryville, Arkansas

## VITA

Philip Matthew Summers is currently the Middle School Principal for Berryville Public Schools in Berryville, Arkansas. Teaching experiences have included grades 7 -12 social studies and college level education courses. He has also served the district as the boys head basketball coach, the Assistant Principal for the high school, and the Transportation Director for the district. Specific areas of interest are educational leadership and curriculum and instruction.

Educational studies have resulted in an Education Specialist Degree in educational leadership from Lindenwood University, a Master of Education in educational leadership from Lindenwood University, and a Bachelor of Science in Education from Southwest Baptist University.

Matt lives in Berryville with his wife and two teenage sons. He is an avid outdoorsman who enjoys fishing, kayaking, and shooting sports. He writes an outdoor column for the Carroll County Newspaper and has filmed two television shows for the Arkansas Game and Fish Commission.