

Lindenwood University

Digital Commons@Lindenwood University

Dissertations

Theses & Dissertations

Spring 5-2009

The STAR Math Test as a Predictor of Arkansas Benchmark Test Scores

Patricia Anne Conner
Lindenwood University

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



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

Recommended Citation

Conner, Patricia Anne, "The STAR Math Test as a Predictor of Arkansas Benchmark Test Scores" (2009).
Dissertations. 549.

<https://digitalcommons.lindenwood.edu/dissertations/549>

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.

Running head: THE STAR MATH TEST AS A PREDICTOR

The STAR Math Test as a Predictor of
Arkansas Benchmark Test Scores

Patricia Anne Conner

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 here at Lindenwood University and that I have not submitted it for any other college or degree here or elsewhere.


Full Legal Name: Patricia Anne Conner

Signature: Patricia Anne Conner Date: 08-10-09

THE STAR MATH TEST AS A PREDICTOR OF
ARKANSAS BENCHMARK TEST SCORES

Patricia Anne Conner

This 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



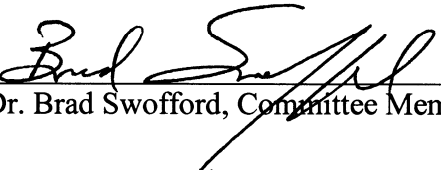
Dr. Terry Reid, Dissertation Chair

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, readership, and valuable insight throughout the entire dissertation process: Dr. Terry Reid, Dr. Sherry DeVore, Dr. Brad Swofford, Mrs. Kathy Grover, and Mrs. Gina Wood.

A special thank you to my family: to my husband Dale who has supported me every step of the way; to my son Justin who I have watched become a father that would have made his grandfather proud.; to my daughter Kelsey who I hope sees a woman can do anything she desires, even if sometimes it must be done backwards and in heels; and to my grandson Ethan who made me realize how much I enjoy the life I have.

Abstract

With the failure of the legislative branch to reauthorize the No Child Left Behind law in 2007 and recent reports that more schools than ever are failing to achieve Adequate Yearly Progress, educators are reviewing practices and curriculum. As a result of federal and state laws, it is necessary to identify an accurate predictor of student achievement prior to the administration of the state-mandated test. For this study, student samples were drawn from sixth, seventh, and eighth grade populations of a Northwest Arkansas Middle School. Samples were separated by grade level and ranked according to the grade equivalency on the fall STAR Math pre-test and the scores on the spring Arkansas Benchmark Test. A quasi-experimental design was implemented to test both the magnitude and reliability of the independent variable, the STAR Math test, on the dependent variable, the Arkansas Benchmark Test. A Pearson r correlation was calculated in each grade level over a three-year period for the relationship between the STAR Math and Arkansas Benchmark. A strong positive correlation was found between the ordinal ranks of grade equivalence on the STAR Math pre-test and the ordinal ranks of the averaged raw score percent on the Arkansas Benchmark Test. Furthermore, a coefficient of determination, a line of best fit, an analysis of variance (ANOVA), and an Omega-squared were used to determine the statistical significance and develop a triangulation of data. Further study is recommended to predict a specific benchmark score based on a STAR Math grade equivalency.

TABLE OF CONTENTS

LIST OF TABLES.....	ix
LIST OF FIGURES.....	xi
KEY TO ABBREVIATIONS.....	xiv
CHAPTER ONE – INTRODUCTION.....	1
Background	1
Conceptual Underpinnings	3
Statement of the Problem	7
Purpose of the Study	8
Questions.....	8
Independent Variable.....	9
Star Math Scores.....	9
Dependent Variable.....	9
Arkansas Benchmark Scores.....	9
Hypotheses.....	9
Null Hypothesis.....	9
Alternative Hypothesis.....	9
Limitations of the Study.....	9
Definitions of Terms.....	12
Summary.....	16
CHAPTER TWO – LITERATURE REVIEW.....	17
Background	17
History.....	17

Implementation of No Child Left Behind.....	19
Accountability.....	24
Systems.....	24
Models.....	32
Assessment Theory.....	38
Assessment for Learning	41
Assessment of Learning	42
Informal Assessment	42
Formal Assessment.....	45
Pre-assessments.....	52
Consequences of State-Mandated Testing.....	56
Positive Consequences	57
Negative Consequences	59
Unintended Consequences	63
Summary.....	65
CHAPTER THREE – DESIGN AND METHODOLOGY.....	67
Introduction.....	67
Population.....	67
Sampling Procedure.....	68
Research Setting.....	70
Research Design	70
Questions.....	70
Independent Variable.....	71

Star Math Scores.....	71
Dependent Variable.....	71
Arkansas Benchmark Scores.....	71
Hypotheses.....	71
Null Hypothesis.....	71
Alternative Hypothesis.....	71
Timeline.....	71
Strategies Applied in the Study.....	74
Coefficient of Determination.....	74
Descriptive Statistics.....	75
Frequency Histogram.....	75
Mean.....	75
Standard Deviation.....	75
Linear Regression.....	76
Omega Squared.....	76
One Way ANOVA.....	76
Pearson Correlation Coefficient.....	77
Survey.....	78
Statistical Treatment of Data.....	78
Magnitude of the Relationship.....	78
Reliability of the Relationship.....	79
Alpha-level.....	79
Ethical and Political Considerations of the Study.....	80

Summary.....	80
CHAPTER FOUR – RESULTS.....	81
Introduction.....	81
Results.....	82
Analysis of Data.....	82
Research Question Number One.....	82
Internal Reliability of the Independent Variable.....	83
Research Question Number Two.....	84
Internal Reliability of the Dependent Variable	84
Research Question Number Three.....	86
Descriptive Statistics of the Raw Data.....	86
Descriptive Statistics of the Ordinal Data.....	90
Linear Regression Results.....	94
Correlation Analyses.....	100
Coefficient of Determination.....	105
Analysis of Variance.....	109
Omega-Squared Results.....	119
Predictive Abilities.....	120
Research Question Number Four.....	121
Survey.....	121
Deductive Conclusions.....	123
Summary.....	123

CHAPTER FIVE – DISCUSSION.....	125
Introduction.....	125
Implication for Effective Schools.....	126
Recommendations.....	127
Summary.....	128
REFERENCES.....	129
APPENDIX A.....	142
APPENDIX B.....	151
APPENDIX C.....	160
E-mail to Arkansas Administrators.....	160
Survey to Arkansas Administrators.....	161
VITA.....	162

LIST OF TABLES

Table 1. <i>Demographics: Study Site School</i>	68
Table 2. <i>Timeline of the Study</i>	72
Table 3. <i>STAR Math Correlations of Pre and Post Tests</i>	84
Table 4. <i>Arkansas Benchmark Correlations</i>	86
Table 5. <i>Comparison of Sample Sizes and Means of STAR Math Tests and Benchmark Tests 2005-2006</i>	88
Table 6. <i>Comparison of Sample Sizes and Means of STAR Math Tests and Benchmark Tests 2006-2007</i>	89
Table 7. <i>Comparison of Sample Sizes and Means of STAR Math Tests and Benchmark Tests for 2007-2008</i>	90
Table 8. <i>Comparison of Sample Sizes and Means of Ordinal Ranks of STAR Math Tests and Ordinal Ranks of Benchmark Tests 2005-2006</i>	92
Table 9. <i>Comparison of Sample Sizes and Means of Ordinal Ranks of STAR Math Tests and Ordinal Ranks of Benchmark Tests 2006-2007</i>	93
Table 10. <i>Comparison of Sample Sizes and Means of Ordinal Ranks of STAR Math Tests and Ordinal Ranks of Benchmark Tests 2007-2008</i>	94
Table 11. <i>Correlations for STAR Math Pre-test Fall and Benchmark Test Spring 2005-2006</i>	102
Table 12. <i>Correlations for STAR Math Pre-test Fall and Benchmark Test Spring 2006-2007</i>	103
Table 13. <i>Correlations for STAR Math Pre-test Fall and Benchmark Test Spring 2007-2008</i>	104
Table 14. <i>Coefficient of Determination between the STAR Math Pre-test and the Arkansas Benchmark Test Spring 2005-2006</i>	106
Table 15. <i>Coefficient of Determination between the STAR Math Pre-test and the Arkansas Benchmark Test Spring 2006-2007</i>	107
Table 16. <i>Coefficient of Determination between the STAR Math Pre-test and the Arkansas Benchmark Test Spring 2007-2008</i>	108

Table 17. <i>One-Way ANOVA Test Sixth Grade 2005-2006</i>	110
Table 18. <i>One-Way ANOVA Test Seventh Grade 2005-2006</i>	111
Table 19. <i>One-Way ANOVA Test Eighth Grade 2005-2006</i>	112
Table 20. <i>One-Way ANOVA Test Sixth Grade 2006-2007</i>	113
Table 21. <i>One-Way ANOVA Test Seventh Grade 2006-2007</i>	114
Table 22. <i>One-Way ANOVA Test Eighth Grade 2006-2007</i>	115
Table 23. <i>One-Way ANOVA Test Sixth Grade 2007-2008</i>	116
Table 24. <i>One-Way ANOVA Test Seventh Grade 2007-2008</i>	117
Table 25. <i>One-Way ANOVA Test Eighth Grade 2007-2008</i>	118
Table 26. <i>Omega-Squared Results</i>	119
Table 27. <i>Survey Results</i>	122

LIST OF FIGURES

<i>Figure 1.</i> Curve Estimation for the Independent Variable and the Dependent Variable for the Sixth Grade 2005-2006.....	96
<i>Figure 2.</i> Curve Estimation for the Independent Variable and the Dependent Variable for the Seventh Grade 2005-2006.....	96
<i>Figure 3.</i> Curve Estimation for the Independent Variable and the Dependent Variable for the Eighth Grade 2005-2006.....	97
<i>Figure 4.</i> Curve Estimation for the Independent Variable and the Dependent Variable for the Sixth Grade 2006-2007.....	97
<i>Figure 5.</i> Curve Estimation for the Independent Variable and the Dependent Variable for the Seventh Grade 2006-2007.....	98
<i>Figure 6.</i> Curve Estimation for the Independent Variable and the Dependent Variable for the Eighth Grade 2006-2007.....	98
<i>Figure 7.</i> Curve Estimation for the Independent Variable and the Dependent Variable for the Sixth Grade 2007-2008.....	99
<i>Figure 8.</i> Curve Estimation for the Independent Variable and the Dependent Variable for the Seventh grade 2007-2008.....	99
<i>Figure 9.</i> Curve Estimation for the Independent Variable and the Dependent Variable for the Eighth Grade 2007-2008.....	100
<i>Figure A1.</i> Frequency Distributions of Grade Equivalency on the Sixth Grade STAR Math Pre-test 2005.....	142
<i>Figure A2.</i> Frequency Distributions of Raw Score Percent on the Sixth Grade Benchmark Test 2006.....	142
<i>Figure A3.</i> Frequency Distributions of Grade Equivalency on the Seventh Grade STAR Math Pre-test 2005.....	143
<i>Figure A4.</i> Frequency Distributions of Raw Score Percent on the Seventh Grade Benchmark Test 2006.....	143
<i>Figure A5.</i> Frequency Distributions of Grade Equivalency on the Eighth Grade STAR Math Pre-test 2005.....	144
<i>Figure A6.</i> Frequency Distributions of Raw Score Percent on the Eighth Grade Benchmark Test 2006.....	144

<i>Figure A7.</i> Frequency Distributions of Grade Equivalency on the Sixth Grade STAR Math Pre-test 2006.....	145
<i>Figure A8.</i> Frequency Distributions of Raw Score Percent on the Sixth Grade Benchmark Test 2007.....	145
<i>Figure A9.</i> Frequency Distributions of Grade Equivalency on the Seventh Grade STAR Math Pre-test 2006.....	146
<i>Figure A10.</i> Frequency Distributions of Raw Score Percent on the Seventh Grade Benchmark Test 2007.....	146
<i>Figure A11.</i> Frequency Distributions of Grade Equivalency on the Eighth Grade STAR Math Pre-test 2006.....	147
<i>Figure A12.</i> Frequency Distributions of Raw Score Percent on the Eighth Grade Benchmark Test 2007.....	147
<i>Figure A13.</i> Frequency Distributions of Grade Equivalency on the Sixth Grade STAR Math Pre-test 2007.....	148
<i>Figure A14.</i> Frequency Distributions of Raw Score Percent on the Sixth Grade Benchmark Test 2008.....	148
<i>Figure A15.</i> Frequency Distributions of Grade Equivalency on the Seventh Grade STAR Math Pre-test 2007.....	149
<i>Figure A16.</i> Frequency Distributions of Raw Score Percent on the Seventh Grade Benchmark Test 2008.....	149
<i>Figure A17.</i> Frequency Distributions of Grade Equivalency on the Eighth Grade STAR Math Pre-test 2007.....	150
<i>Figure A18.</i> Frequency Distributions of Raw Score Percent on the Eighth Grade Benchmark Test 2008.....	150
<i>Figure B1.</i> Frequency Distributions of the Ordinal Ranks for Grade Equivalency on the Sixth Grade STAR Math Pre-test 2005.....	151
<i>Figure B2.</i> Frequency Distributions of the Ordinal Ranks for Raw Score Percent on the Sixth Grade Benchmark Test 2006.....	151
<i>Figure B3.</i> Frequency Distributions of the Ordinal Ranks for Grade Equivalency on the Seventh Grade STAR Math Pre-test 2005.....	152

<i>Figure B4.</i> Frequency Distributions of the Ordinal Ranks for Raw Score Percent on the Seventh Grade Benchmark Test 2006.....	152
<i>Figure B5.</i> Frequency Distributions of the Ordinal Ranks for Grade Equivalency on the Eighth Grade STAR Math Pre-test 2005.....	153
<i>Figure B6.</i> Frequency Distributions of the Ordinal Ranks for Raw Score Percent on the Eighth Grade Benchmark Test 2006.....	153
<i>Figure B7.</i> Frequency Distributions of the Ordinal Ranks for Grade Equivalency on the Sixth Grade STAR Math Pre-test 2006.....	154
<i>Figure B8.</i> Frequency Distributions of the Ordinal Ranks for Raw Score Percent on the Sixth Grade Benchmark Test 2007.....	154
<i>Figure B9.</i> Frequency Distributions of the Ordinal Ranks for Grade Equivalency on the Seventh Grade STAR Math Pre-test 2006.....	155
<i>Figure B10.</i> Frequency Distributions of the Ordinal Ranks for Raw Score Percent on the Seventh Grade Benchmark Test 2007.....	155
<i>Figure B11.</i> Frequency Distributions of the Ordinal Ranks for Grade Equivalency on the Eighth Grade STAR Math Pre-test 2006.....	156
<i>Figure B12.</i> Frequency Distributions of the Ordinal Ranks for Raw Score Percent on the Eighth Grade Benchmark Test 2007.....	156
<i>Figure B13.</i> Frequency Distributions of the Ordinal Ranks for Grade Equivalency on the Sixth Grade STAR Math Pre-test 2007.....	157
<i>Figure B14.</i> Frequency Distributions of the Ordinal Ranks for Raw Score Percent on the Sixth Grade Benchmark Test 2008.....	157
<i>Figure B15.</i> Frequency Distributions of the Ordinal Ranks for Grade Equivalency on the Seventh Grade STAR Math Pre-test 2007.....	158
<i>Figure B16.</i> Frequency Distributions of the Ordinal Ranks for Raw Score Percent on the Seventh Grade Benchmark Test 2008.....	158
<i>Figure B17.</i> Frequency Distributions of the Ordinal Ranks for Grade Equivalency on the Eighth Grade STAR Math Pre-test 2007.....	159
<i>Figure B18.</i> Frequency Distributions of the Ordinal Ranks for Raw Score Percent on the Eighth Grade Benchmark Test 2008.....	159

KEY TO ABBREVIATIONS AND SYMBOLS

AIP	Academic Improvement Plan
AYP	Adequate Yearly Progress
α	Alpha level
H_1	Alternative Hypothesis
ANOVA	Analysis of Variance
ACTAAP	Arkansas Comprehensive Testing and Accountability Program
ADE	Arkansas Department of Education
r^2	Coefficient of Determination
EOC	End-of-Course Exams
GE	Grade Equivalency
ITBS	Iowa Test of Basic Skills
NCLB	No Child Left Behind
H_0	Null Hypothesis
ω^2	Omega-Squared
r	Pearson r Correlation Coefficient
SAT	Scholastic Achievement Test
SD	Standard Deviation
p -value	Statistical Significance

CHAPTER ONE – INTRODUCTION

Background

The failure of the United States legislature to reauthorize the No Child Left Behind (NCLB) Act in 2007 left its future in question (Klein, 2008). However, with the appointment of Arnie Duncan as President Obama's Secretary of Education, it is clear that NCLB is not going away. Duncan reported the House Education and Labor Committee, "The No Child Left Behind Act, with a focus on accountability, was a huge step in the right direction" (Hoff, 2008d, p. 25). Even so, approximately 30,000 United States public schools failed to meet Adequate Yearly Progress (AYP) in the 2007-2008 school year (Hoff, 2009), and there is still a concern the achievement gap for poor and minority students is not being closed (Wilcox, 2007).

With new direction for NCLB on the horizon and the failure of schools to meet targets, especially for poor and minority students (Cavanaugh, 2008), there has been a renewed effort by educators to review practices and curricula. Resources are being redirected in an effort to predict and enhance student achievement (Clark, Madaus, Ramos, Lynch, & Lynch, 2001). The Herculean task is to overcome the inherent flaws in the current NCLB, and provide an effective economically efficient means to achieve AYP, while maintaining an enriched curriculum.

After the passage of the original NCLB mandate, states were left to implement the law with few guidelines (United States Department of Education [DOE], 2001). In 1998, Arkansas developed a criterion-referenced test in literacy and math and began a

field study of the achievement of fourth grade students. Since that time, grade levels three through eight were required to assess state benchmarks. At the high school level, End-of-Course (EOC) exams were compulsory in Algebra I, Geometry, and eleventh grade literacy (Arkansas Department of Education [ADE], 2004). Science was added in 2006 to the fifth and seventh grade levels, and then EOC tests in Biology and Algebra II were implemented in the spring of 2007 (ADE, 2007a). Furthermore, each year, a norm-referenced test, the Iowa Test of Basic Skills (ITBS), was given to all Arkansas students, kindergarten through grade nine (ADE, 2004).

However, in 2007, Arkansas contracted with Harcourt Pearson to develop an Augmented Benchmark which included both criterion and norm-referenced questions. The norm-referenced portion of the test became Harcourt Pearson's Standardized Achievement Test (SAT 10). The contract with this company will run through the year 2013 (Gray, 2007). As a result of the change, it will be at least three years before any real student achievement trends are identifiable. Harcourt Pearson released a correlation chart between the previous ITBS and the current SAT 10 in order to develop a baseline for comparison (ADE, 2008b).

In the rules governing the Arkansas Comprehensive Testing, Assessment, and Accountability Program (ACTAAP) handed down by the Arkansas Department of Education (ADE, 2004), the criteria for meeting the mandated AYP were detailed. These rules were designed to achieve the following: clear academic standards that are periodically reviewed and revised; professional development standards for all administrators, teachers, and instructional support personnel; expected achievement levels; reports on student achievement and other indicators; school and school district

evaluation data; a system of sanctions and rewards based on performance of schools and school districts; and compliance with current federal and state law and State Board of Education policies (ADE, 2004). Furthermore, Arkansas law required that each school district create an Academic Improvement Plan (AIP) for each student who did not achieve proficient or advanced on the state-mandated test. This plan must detail how remediation will be achieved in all deficient areas (ADE, n.d.). As a result, Arkansas schools are studying efficient methods of predicting proficiency rates and providing interventions (ADE, 2004).

Conceptual Underpinnings

Whether or not a district embraces testing is of little consequence. Federal mandates have required states to establish a testing process (DOE, 2001), and districts must comply or suffer funding losses (ADE, 2004). In order to abide by these laws and still serve the public's welfare, scarce resources must be used wisely. Educational decisions involve delegating resources, such as time and money, in ways which will increase student achievement (Miles, 2001). It is vital schools base decisions on evidence.

In a desire to predetermine student achievement levels, schools have been creating or purchasing formative assessment systems to monitor student progress (Chappius & Chappius, 2008). These assessments provide essential data to raise achievement levels. Results of formative assessments are used to evaluate and plan instructional practices (Marzano, 2006; Ravitch, 2007). While state tests are important summative communicators of student proficiency (Ravitch, 2007) and allow schools to reform curriculum and instruction long-term, the tests do not provide on-going information

that schools may employ to incrementally improve instructional programs (Popham, 2007b). Furthermore, the tests do not address the learning problems of students with the most need (Herman & Baker, 2005).

Formative assessments have taken on an essential role in the current educational reform environment. The idea, however, was not a new one. Black and Wiliam in a 1998 article, “Inside the Black Box: Raising Standards through Classroom Assessment,” set the stage for worldwide interest in formative assessment (Popham, 2007b). Black and Wiliam (1998) stated:

Present policies in the U.S. and in many other countries seem to treat the classroom as a black box. Certain *inputs* from the outside – pupils, teachers, other resources, management rules and requirements, parental anxieties, standards, tests with high stakes, and so on – are fed into the box. Some *outputs* are supposed to follow: pupils who are more knowledgeable and competent, better test results, teachers who are reasonably satisfied, and so on. But what is happening inside the black box? How can anyone be sure that a particular set of new inputs will produce better outputs if we don’t at least study what happens inside? And why is it that most of the reform initiatives are not aimed at giving direct support to the work of teachers in classroom? (p. 139)

The research in the Black and Wiliam (1998) study presented a meta-analysis of 23 studies which concentrated on classroom assessment and incorporated a significant number of innovations. The results concluded the practices analyzed presented substantial learning gains. The effect sizes of the formative assessment experiments

were between 0.4 and 0.7 which translated into percentile gains of 35 or higher (Black & Wiliam). The effect size represented a standardized measure of the effect of an intervention on student outcomes (DOE, 2008), and “the effect sizes for summative assessment are consistently lower than the effect sizes for formative assessments” (Marzano, 2006, p. 9).

A more recent case study was performed by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). According to Herman and Choi (2008), the study sampled thirteen middle school teachers with seven teachers completing the unit and all other data requirements. While the authors admitted the small sample size did not lend the results to any firm empirical base, they did believe the results were promising. Herman and Choi concluded that formative assessment enabled teachers to know where students were performing relative to learning goals, despite a small sample and imperfect reliability measures. The results showed the more accurate teachers were in the knowledge of where students were, the more effective teachers would be in promoting subsequent learning (Herman & Choi).

The question is: how do educators create a symbiotic relationship between formative and summative assessments? Furthermore, why would districts want to? In an ideal assessment system, both formative assessment and cognitive learning work together to inform teaching and improve performance (Baker, Herman, & Linn, 2005; Marzano, 2006; Popham, 2007a). Currently, states have been wrestling with finding the right balance for local formative assessments with a statewide assessment to be used for state accountability purposes. Lewis (2005) believed that, “Appropriately designed assessment situations can have substantial impact on the

quality of information provided to teachers and students for instructional decision-making and meaningful learning” (p. 5).

The states goal was to design a policy to accommodate assessment for learning though the use of formative assessments that provided timely and informative feedback to improve instruction on a regular basis and assessment of learning to ensure that all students were seeing increased student achievement (Conrad, 2008). Popham (2007b), a well-known critic of NCLB, believed that formative assessment has the potential to aid student achievement and help districts reach AYP. Popham concluded:

If formative assessment improves student learning in the classroom, couldn't it *also* improve test scores on external accountability tests? Considering that so many educators are now figuratively drowning in an ocean of accountability, it not surprising to see formative assessment cast in the role of life preserver. If it is true that drowning people will grasp at straws in an effort to stay afloat, it is surely as true that they will grasp even more eagerly at “research proven” straws. (p. 5)

A new wealth of immediate student data that has been provided by formative assessments presented educators with a conduit for decision-making (Marzano, 2006). These assessments serve a variety of purposes, including predicting a student's ability to succeed on a large-scale summative assessment, evaluating a particular educational program, or diagnosing gaps in student's learning (Perie, Marion, & Gong, 2007). Olson (2007) stated, “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” (p. 11).

Statement of the Problem

In essence, what can a set of test scores tell about the quality of education and the relationship to student performance? In an attempt to respond to this question, the overarching problem emerged: There is a need to identify a statistically significant predictor of student achievement that can be monitored over time and used as a source for remediation and early intervention. One available assessment is the STAR Math test. However, since program costs are considerable, it is essential that a district weigh the cost effectiveness against desired student achievement outcomes (Miles, 2001). STAR Math was developed by the Renaissance Learning Company (2006) and offers computer-adaptive tests which provide the respondent with a grade equivalency and percentile math scores for grades first through twelfth in less than fifteen minutes. The accompanying Accelerated Math Program supports curriculum by providing individualized practice tailored to each student’s weakness, immediate results, and continuous feedback. All of the math questions are linked to recommendations provided by the National Council on Teaching Mathematics (NCTM). This was an important factor as the Arkansas student learning expectations were also closely aligned to the NCTM (Renaissance Learning Company, 2006).

The brochure about the STAR Math test published by the Renaissance Learning Company stated (2006), “teachers can predict achievement on state standards, determine the appropriate level of challenge, instantly place new students, identify those who need individual help, and plan individualized instruction on the skills

students need to become successful at math” (p. 7). A study to determine whether the STAR Math predicted student achievement on the Arkansas Benchmark test would allow a district to make more informed financial decisions. Furthermore, if a specific STAR Math grade equivalency was determined to have a high correlation of success in predicting a proficient Benchmark score the information would be invaluable.

Purpose of the Study

Arkansas school districts are challenged to achieve AYP as state funding has been directly linked to test scores (ADE, 2004). Therefore, the purpose of the study was to identify an assessment to accurately predict student achievement, target student weakness prior to the benchmark test, and focus efforts on direct remediation. If target areas are identifiable, districts can restructure curricula more effectively and efficiently. Intelligent fiscal policy is imperative and some educators have advocated a reduction of spending on non-academic teaching staff (Miles, 2001). However, by using scarce resources wisely, it is not necessary to find a quick fix by shifting dollars away from programs such as fine arts or vocational classes. Budgets can be planned more successfully while meeting students’ needs and maintaining AYP.

Questions

The following questions were addressed in the study:

1. What relationship exists between the STAR Math pre-test and post-test in the sixth, seventh, and eighth grades for 2006, 2007, and 2008?
2. What relationship exists between two consecutive years of the Arkansas Benchmark in the sixth, seventh, and eighth grades for 2006, 2007, and 2008 using corresponding student populations?

3. What relationship exists between the STAR Math pre-test and the Arkansas Benchmark examination in the sixth, seventh, and eighth grades from 2006, 2007, and 2008?
4. How do Arkansas administrators view the use of pre-assessments as an indicator of achievement on the Arkansas Benchmark Test?

Independent Variable

The independent variable in the study was the Star Math pre-test scores.

Dependent Variable

The dependent variable in the study was the Arkansas Benchmark Test scores of corresponding students.

Hypotheses

Null Hypothesis

There is no significant relationship between the STAR Math Test and student achievement on the Arkansas Benchmark Test. The null hypothesis is designated by the symbol H_0 .

Alternative Hypothesis

There is a significant relationship between the STAR Math Test and student achievement on the Arkansas Benchmark Test. The alternative hypothesis is designated by the symbol H_1 .

Limitations of the Study

Ceiling and floor effects. Ceiling and floor effects make it difficult to distinguish the higher and lower ends of a normal curve distribution because the starting and stopping points do not allow for movement any farther up or down than the finite

scale allows (Ravitch, 2007). This limitation was unavoidable since the STAR Math uses a grade-equivalency scale from 1 to 12.9 and the Arkansas Benchmark raw score percents were from 1 to 100. The primary interruption of the normal curve distribution in this study occurred at the 12.9 grade equivalency on the STAR Math test data and the number one ordinal rank position on the STAR Math rankings.

Factors beyond the scope of this study. There were important uncontrollable factors to student achievement such as teacher quality, curriculum quality, parental involvement, socio-economic status, and language barriers. Efforts to minimize these factors included limiting the student sample population. Only students who participated in the previous year's benchmark, the STAR Math pre- and post-test for the current year, and the Arkansas Benchmark test for the measured years were included. This provided a small degree of consistency for teacher and curriculum standards.

Maturation. This study was limited by the emergence of personal and behavioral characteristics through growth processes, or maturation (2009). Even though the study spanned three years, efforts to minimize maturation were made by separating the data sets and examining the statistical relationships by individual years and individual grades.

Predictive ability. A limitation on a test such as the STAR Math Test happens when the assessment provides quality feedback on student learning improvements, then the predictive ability is likely to decrease. According to Perie, Marion, and Gong (2007):

If the test predicts that a student is on-track to perform at the basic level, and then appropriate interventions are used to bring the student to proficient, the statistical analysis of the test's predictive validity should under predict student performance over time. (p. 17)

Research design. The effort to minimize uncontrollable factors such as teacher quality and curriculum quality made a random sampling procedure impossible. As a result, a quasi-experimental design was implemented which does not apply a random sample but used a series of multiple measures over multiple years (Trochim, 2008).

School participation. This study involved one Northwest Arkansas School District. Future studies would benefit by increasing the number of the participating schools.

Survey design and response. The survey was created by the researcher. Two hundred surveys were sent via e-mail and 92 educators responded. The researcher assumed that all respondents answered each question honestly.

Test development and administration. The test developer changed during the third year of the study, and proficiency scales were also adjusted. To minimize the effect of the change, a measurement of reliability was conducted. A Pearson r correlation coefficient was calculated between a test from one year to the next and included only the scores from students who participated on both tests. To overcome the latter problem scale scores were converted to raw scores and percents were calculated. The Arkansas Benchmark Test is a standardized criterion-referenced assessment. Human error during administration is always a

potential limitation; however, to decrease the likelihood of mistakes, all staff members were trained and followed test security guidelines provided by the test manufacturer.

Definitions of Terms

Academic Improvement Plan (AIP). This is a plan detailing supplemental or intervention and remedial instruction, or both, in deficient academic areas for any student who is not proficient on a portion or portions of the state-mandated criterion-referenced assessments. A student's failure to remediate can result in his/her retention (ADE, 2004).

Adequate Yearly Progress (AYP). An individual state's measure of yearly progress toward achieving state academic standards, as described in the NCLB legislation (DOE, 2001). AYP is the minimum level of improvement that states, school districts, and schools must achieve each year (ADE, 2004; Ravitch, 2007).

Advanced/Proficient. An achievement score which is calculated by a percent of the raw score on a criterion-referenced test determined by the state as necessary to meet AYP (ADE, 2004). Two of the three achievement levels on the federally funded National Assessment of Educational Progress (NAEP) test (Ravitch, 2007).

Alternate Portfolio. An alternative assessment method used in Arkansas to assess achievement of students who can not otherwise take the Arkansas Benchmark examination due to severe cognitive disabilities (ADE, 2004).

Arkansas Comprehensive Testing, Assessment and Accountability Program (ACTAAP). This is a comprehensive system that concentrates on high academic standards, professional development, student assessments, and accountability for all schools. 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 (Figlio, 2008).

Ceiling effect. This is the tendency of students at the top of the achievement scale not to increase their test scores dramatically because they have already reached the ceiling, or the highest possible level of achievement. When scores in a high-performing school remain stagnant, it may be because there is relatively little room for improvement and virtually no room for large gains on the kinds of assessments being used (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 (Bond, 1996; Ravitch, 2007).

Data-driven decision-making. This term refers to teachers, principals, and administrators systematically collecting and analyzing various types of data, including input, process, outcome, and satisfaction data, to guide a range of decisions to help improve the success of students and schools (Pascopella, 2006).

Floor effect. These are items on a norm-referenced test that are too hard to discriminate at the lower end of the ability scale. This is the lowest level of a performance that is measured by a test (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). Also defined as any activity that provides sound feedback on students' learning (Marzano, 2006).

Growth model. This is a model that provides a method for tracking student progress over a period of time (Goldschmidt & Choi, 2007).

Informal assessment. This assessment collects data by any other means than a standardized test (Rabinowitz, 2005).

Interim assessment. These assessments are designed to measure progress during a course of instruction, usually administered periodically throughout the year to monitor student progress at meeting state standards (Perie, et al., 2007; Ravitch, 2007).

No Child Left Behind Act. A legislative act initiated by the George W. 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; Ravitch, 2007).

Norm-Referenced Tests. An assessment designed to compare the scores of individuals or groups of individuals with the scores achieved by a representative sample of individuals with similar characteristics, members of a so-called reference group. Norm-referenced tests are useful for comparing the performance of students in one school, district, country, state or nation with the performance of students in others (Bond, 1996; Ravitch, 2007).

Pre-assessment. This is an assessment designed to discover what students know prior to the instruction so that curriculum and practices are driven by this knowledge (Popham, 2007b).

Report card. Under NCLB, states must require districts to publicly report state-mandated assessment information and provide explicit information to students, parents and teachers about the results of student progress (Crone, 2004).

School choice. Schools that do not meet Adequate Yearly Progress must inform parents of the right to withdraw their children from the district and place them in a higher performing school without penalty (ADE, 2004).

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

STAR Math Test. This is a computerized math test developed by the Renaissance Learning Company. This program places students at the appropriate instructional grade level and provides remediation practices through the Accelerated Math Program (Renaissance Learning, 2006).

Student achievement. This is 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 activities (Ravitch, 2007).

Value-added assessment. A method of gauging the effect that a school, a teacher, or a program has on student learning by measuring and comparing the gains in student performance over time. The difference between the measures represents the learning gain (Ravitch, 2007).

Summary

The original NCLB Act required each state to develop a standards-based criterion-referenced benchmark test and to establish a definition of AYP. Since federal funding has been tied to AYP (DOE, 2001), public schools examine every available option in order to meet these mandated goals. This study was designed to examine whether the STAR Math test is an accurate predictor of student achievement on the Arkansas Benchmark Test. The STAR Math test, as a potential predictor of achievement and a resource for remediation of weaknesses of individual students, was worthy of investigation. Schools considering the purchase of any predictive/remediation program should analyze the cost versus the benefits of such an expensive endeavor.

The review of literature in chapter two addressed myriad viewpoints and strategies concerning assessment and accountability systems. A description of the design, methodology, and statistical strategies used to analyze findings was offered in chapter three. Chapter four discussed the results of the research, and chapter five provided implications and recommendations for future study.

CHAPTER TWO – LITERATURE REVIEW

This chapter involved an examination of the No Child Left Behind (NCLB) mandate and its effect on public school accountability. The review analyzed various accountability systems and ensuing models as researchers and lawmakers have agreed a one-shot test is not a true picture of student achievement. It has become necessary to take students' initial achievement levels into account (Weiss, 2008). The different assessment categories and the function of standardized testing, and any potential consequences, intended as well as unintended, associated with testing were explored. An additional factor explored was the specific function of pre-assessment as a potential indicator of student performance on standardized tests, especially on the Arkansas Benchmark Accountability Assessment Test (ACTAAP).

Background

History

Government involvement in education has been common-place, from the passage of the Massachusetts Old Deluder Law in 1647 to the present (Crone, 2004). Since the nation's inception, the General Welfare Clause in the United States Constitution provided government with the necessary means of participation in education. However, early assessments were not dictated by government sanction. They were informal, primarily teacher-made tests which were certainly not lacking in rigor (Crone, 2004).

The development of the Stanford Achievement Test (SAT) in 1923 allowed for standardized testing which opened the door for government involvement into education. This attachment has increased over time. The military first used standardized testing for placement purposes. Between 1941 and 1960 these formal assessments held students and curricula accountable; not public schools (Crone, 2004). It would be 2001 before standardized tests became the meter by which public schools were judged (DOE, 2001).

The Elementary and Secondary Education Act implemented in 1965 and established under President Johnson's Great Society opened doors for various Title programs which are still in existence today (Crone, 2004; Popham, 2007a). These Title funds were directed toward impoverished students and testing has become a means to appraise the corresponding program's effectiveness. One such test used for evaluation purposes is 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 is designed to measure progress over time (Crone, 2004). NAEP's current application assists in the diagnosis of a state's testing programs as it tries to comply with NCLB.

However, it was the 1983 report by the National Commission on Excellence in Education that spotlighted nation-wide attention on public schools. The now famous or infamous study, *A Nation at Risk: The Imperative for Education Reform*, asserted that the national education system was in complete disarray (Wong & Nicotera, 2007). Furthermore, the report stated such was the status of education that it compromised the country's preeminence, both technologically and militarily (Wong

& Nicotera, 2007). This played directly on the fears highlighted by events of the cold war, and these fears were heightened by President Regan who used this as a stand against communist world domination (Crone, 2004).

Even though it was later believed that the report made exaggerated claims about the decline of student achievement, the questions it raised were not put to rest (Wong & Nicotera, 2007). However, despite the obvious research shortfalls another inherent deficiency was identified: the report did not put into place an accountability system to carry out the recommendations offered within the study (Wong & Nicotera, 2007). It was nearly twenty years before the NCLB legislation enacted a federal accountability system. This mandate signed into law in 2002 emphasizes high stakes testing in a manner that is changing the face of education (ADE, 2004).

Implementation of No Child Left Behind

By the 2005-2006 school year, all students in grades three through eight were tested annually in math and literacy. States developed, administered tests, and specified what constituted an allowable proficiency rating for each grade. This flexibility permissible in the legislation caused groups such as the National Association of Secondary School Principals (NASSP) to voice concerns (Kennedy-Manzo, 2008). In a position statement, NASSP asked Congress to create an independent panel of researchers and educators to develop common guidelines for proficiency in mathematics and literacy. NASSP stated, “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).

In Arkansas, four factors contribute to a school's Adequate Yearly Progress (AYP) and determine whether or not a district is placed on the school improvement list. The first factor is a student assessment in both mathematics and literacy (ADE, 2007a). This is a criterion-reference test aligned to state standards at each grade level three through eight. There are also End-of-Course Exams for Algebra I, Geometry, Algebra II, and Biology as well as Eleventh Grade Literacy tests (ADE, 2007a). The second factor necessary to achieve proficiency is the requirement that 95 percent of all eligible students must participate in these academic assessments. The third factor states 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 factor 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. These populations are considered eligible when the total sub-group for a building is 40 or more students (ADE, 2005).

Since NCLB allows states to create or purchase achievement tests, 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 subjected to examination. Due to the fact NAEP is the only standardized test administered to a representative sample of students across the nation, it is often referred to as the Nation's Report Card (Cavanaugh, 2008). Since 1969, NAEP 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 (Cavanaugh, 2008). In 2007, the NAEP writing assessment was administered to approximately 161,000 eighth graders in more than 7,640 schools between January and March. In Arkansas, about 4,900 students in 260 schools took part in the exam (ADE, 2007b). NAEP results were reported both as scores and also as performance levels. The names of these performance levels are similar to those used to report Arkansas benchmarks, though they represented slightly different groupings of students (ADE, 2007a). Dr. Ken James, Commissioner of the Arkansas Department of Education, spoke encouraging words in a September, 2007 *News Release* when discussing the latest NAEP report, "We know we have the right pieces in place to put together a successful learning experience for all of our students...I fully expect that the positive results we have witnessed in recent years will continue" (p. 2).

NCLB mandated all school districts reach 100 percent proficiency of student achievement on state department approved tests by the end of the 2013-2014 school year. As well as designing achievement tests, states were 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. As a result, these indicators varied drastically among the different states (Fuller, Wright,

Gesicki, & Kang, 2007). Analysts predicted that by the 2013-2014 school year, a majority of school districts will not meet AYP requirements, even many of America's highest achieving schools in affluent areas that statistically score above the national achievement mean (Goldschmidt & Choi, 2007). Furthermore, the American Institute for Research reported that two-thirds of state education departments do not have adequate capacity to help low-performing schools (McNeil, 2008).

What does the future hold for NCLB? The fifth anniversary of the federal bipartisan legislation has come and gone, and reauthorization for the mandate was due in 2007. However, reauthorization was delayed as calls for change came from even those who typically supported the legislation; including the conservatives who voted overwhelmingly for the original bill (Klien, 2008). Michael Petrilli, vice president for national programs and policy at the Thomas B. Fordham Foundation, spoke at a conference by the American Enterprise Institute on November 30, 2006. He reiterated the views of the conservative foundation; the intent to close the achievement gap for poor and minority students is failing Petrilli imparted the recommendation that NCLB must be readjusted if it is to remain school improvement leverage (Wilcox, 2007).

Further cause for concern was that one of the chief sponsors of the original bill, Senator Edward Kennedy, who reached across the aisle to aid the passage of NCLB, became ill with a malignant brain tumor (Klein, 2008). Washington legislators have become concerned that without his forceful presence, prospects for reauthorization were grim. Senator Kennedy stated the law should be more flexible than its original form and reward schools for individual student's progress (Klein, 2008). He also

believed the federal government should help struggling schools more by providing additional resources (Klein, 2008).

The Fordham Institute in Washington D.C. surveyed twenty education insiders about their predictions for NCLB prior to its reauthorization date. All but one of the respondents believed that the legislation would not be reauthorized until after the 2008 presidential election, and a majority felt only small adjustments would be made. They also considered that the core of any change should center on a growth model plan which integrated a variety of measures for accountability (Loup & Petrilli, 2005).

November of 2008 brought a presidential election and the two presumptive nominees for the democrat and republican parties had spoken about NCLB. The problem was that neither candidate proposed any new or concrete plans, nor said what he would do about the future of NCLB. They did not address the goal of 100 percent proficiency by the end of 2014 or how to improve interventions in schools not meeting the goals set out in the law (Hoff, 2008a). Hoff reported:

The Democratic campaign advisor stated about his candidate, Barack Obama, He views continuing down our present path as morally unacceptable and economically untenable ... it is time to move beyond the tired debates of the past and towards a new era of reform while his Republican opponent John McCain released an equally impotent statement that he would lead a renaissance in education that would make significant changes to the K-12 system. (p. 24)

According to an *Education Week* article, Reg 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 (Wilcox, 2007). His first suggestion was the use of multiple measures and methods to gauge achievement and school quality to determine school effectiveness. Weaver stated these measures should gauge growth over time and not be solely based on a certain proficiency level (Wilcox, 2007).

Regardless of how educators viewed the mandate, it has been their responsibility to fulfill the policy provisions, and individual states and local districts have been trying to make sense of the law while putting theory into reality. States must take every precaution to create accountability systems which avoid unintended, negative results (Stecher & Hamilton, 2002). The goal must be to meet federal regulations and use reform measures to actually drive curriculum changes thus increasing student achievement.

Accountability

Systems

The public demanded school accountability; the legislative and executive branches of the federal government in rare bi-partisan form mandated school accountability (Goodwin, et al., 2003; Wilcox, 2007); and even educators recognized the necessity for schools to provide quality instruction and increased achievement for all students. Therefore, it was no surprise that this has become one of the prominent parts of NCLB, and schools failing to make adequate progress faced severe consequences, which include reconstitution, state takeover, or closure (Goodwin, Englert, &

Cicchinelli, 2003). This development in the national philosophy encouraged states to reexamine how school districts have been evaluated, and the resulting evolution of accountability systems has been inevitable. Numerous studies on this topic have been initiated to develop guidance for states (Weiss, 2007).

Proponents of a no excuses accountability system believed it is essential to set clear expectations for students and to hold educators responsible for guaranteeing that student achievement is met (Weiss, 2007). Expectations include a focus of schools and districts on learning outcomes and how well students are learning; a focus of teachers on reaching all groups and helping them achieve; including economically disadvantaged, special needs, and limited English students (DOE, 2001).

Additionally, a component of the law required that every classroom be instructed by a highly-qualified teacher. The educator must have the proper credentials to teach the subject to which he or she was assigned (ADE, 2004). Many districts lack highly-qualified teachers; particularly those districts suffering from low socio-economic levels. It has been difficult to draw qualified educators, especially during a teacher shortage, when in a competitive market higher socio-economic communities pay elevated wages and offer better working conditions (Barton, 2006).

Proponents of a stringent accountability system also realize it has become vital to inform parents as to how well children and schools perform. One critical element of NCLB was the school choice provision which allowed parents to leave failing schools (DOE, 2001). In Arkansas, schools failing to meet AYP are placed on year one of School Improvement. There are consequences associated with this label. Parents must be notified in writing about the designation and the fact they may withdraw children

from the district and place them in a higher performing school without penalty. In addition, the penalized school must offer after-school tutoring and initiate programs designed to increase student achievement (ADE, 2007a).

There have been potential negative consequences of strict accountability systems expressed by members of the education community, such as; the potential side effects of unintentionally narrowing the curriculum as teachers teach to the test (Deubel, 2008), and a focus on proficiency levels rather than growth where proficient students are then ignored and not brought to an advanced level (Cech, 2008b). There has been fear that a strong accountability system will result in an increase in retention rates or an increase in the placement of students in special education in an attempt to elevate scores by allowing improper accommodations (Cavanaugh, 2008). Accommodations, such as extended time on a standardized test, benefit students who should be in the regular education program and give them an advantage over students not receiving these modifications (Cavanaugh, 2008).

Arkansas allows for certain student populations to complete testing through an alternate portfolio system. It has been designed to evaluate the performance of students with significant cognitive disabilities (ADE, 2008a). The alternate portfolio must be administered for literacy in grades three through eighth and the eleventh, mathematics in grades three through the eighth, and science for grades five and seven. All ninth grade students with disabilities who have not taken Algebra I or geometry must be assessed with an alternate portfolio for math. Additionally, all tenth grade students with disabilities who have not taken biology must be assessed with an

alternate portfolio for science. However, there is a one percent cap on the number of students who receive this modification (ADE, 2008a).

Prior to the 2007-2008 school year, Arkansas English Language Learners (ELL) students were also eligible to complete an alternate portfolio providing there was a committee designation (ADE, 2004). However, this application was stopped by the federal government when it refused to renew Arkansas' accountability system until the practice was changed. Arkansas was not the only state to struggle with educating non-English students. A nation-wide achievement gap exists between this sub-population and their English speaking counter parts (Zehr, 2008). The 2007 NAEP report stated that fourth grade ELL students tested in reading had only a 7.5 proficiency rating while English speakers had a 35.5 proficiency rating (Zehr, 2008). Consequently,

An unfortunate outcome of all the fine print of the NCLB mandate and the ensuing accountability systems is the potential reaction to focus more on responding to bureaucratic regulations rather than addressing other issues of greater concern, and furthermore, adopt a compliance mentality, rather than a creative improvement mindset. (Goodwin, et al., 2003, p. 3)

A McRel *Policy Brief* (Stapleman, 2000) examined one study which presented six points to consider when developing an accountability system. The study pointed out that 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 was unfair to mandate educators to teach a certain set of content standards, and then 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 (Stapleman, 2000).

The *Brief* points out that in this litigious society the accuracy of these high-stakes consequences will be challenged. Fourth, the accountability system should provide several performance indicators and not hinge on a single test score. Possible variables include student achievement, attendance, drop-out rates, and graduation rates (Stapleman, 2000). This point has been a common theme among the various studies developed on accountability systems. Fifth, there needed 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 (Stapleman, 2000). Kohn (2001b) provided criteria for judging standards. He believed standards should be non-specific and the more specific the standard the further students and teachers are distanced from the learning process (Kohn, 2001b). There was no room for creativity and investigation when the goal simply was to cover massive amounts of material; therefore he didn't believe that standards had to be measurable. Kohn stated, "Measurable outcomes may be the least significant results of learning" (Kohn, 2001b, ¶ 3). Kohn questioned uniform standards where all

students must learn exactly the same thing. Lastly, Kohn believed standards must be considered guidelines rather than mandates.

The second part of the model for standards-based accountability systems developed by the Education Commission of the States (Stapleman, 2000), 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 (Stapleman, 2000). 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 (CREST) also developed criteria for an accountability system. Like the two previous reports, an emphasis was placed on employing different types of data from multiple sources (Baker, Linn, & Herman, 2002). Furthermore, it called for a report card where results have been made available and understandable with all elements in the system explicitly identified (Baker, et al., 2002). A difference in this report from the aforementioned was that it took into account the performance of all students including subgroups that historically have been difficult to assess. In addition, 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). These studies have been essential in order to judge the effectiveness of existing accountability systems by allowing schools to know in advance how the process would operate in practice and the effect it would produce.

Furthermore, these studies also allowed for continued improvement in testing programs and accountability systems (Baker, et al., 2002).

If the aspiration of accountability systems is to increase student achievement, the question that must be asked becomes; how exactly will school accountability lead to this improvement? Unfortunately, this question is rarely addressed or answered (Goodwin, et al., 2003). Similar to the studies mentioned, Wong and Nicotera (2007) called for the establishment of clear goals for academic and performance standards:

When goals throughout the education system are focused on academic and performance standards, teachers will have the capacity to make changes to their instructional practices and increase academic press. Academic press consists of a combination of high-quality homework, course content, and teacher expectations. (p. 28)

There have been certain assumptions about outcome-oriented accountability systems. The belief exists that schools would be improved by publicly reporting assessment information and providing explicit information to students, parents, and teachers about the results of student progress (Crone, 2004). NCLB requires states to publish report cards which describe student performance on standardized tests (DOE, 2001). The danger is the public would view these scores as the foremost measure of school quality (Crone, 2004). As a result, school districts are appraised on the basis of AYP whether there is achievement or not. Arkansas meets this requirement by publishing a school report card on the Arkansas Department of Education (ADE) web-site and sending out the report card with individual student information to each parent (ADE, 2007a). The assumption was that the parents will draw on this

information to demand improvement, and teachers will disaggregate data by subgroups to discover specific strengths and weaknesses and devise plans to help improve student learning.

There was also the conjecture that the learning process is being monitored and that students, teachers, districts, and states are being held accountable for attaining desired learning outcomes (ADE, 2004). The focus is no longer on how the information is being taught, but on what is being taught and how well it is affecting achievement. The goal, hopefully, has become to provide schools with more flexibility with which to maximize student learning.

Another common idea was these accountability systems determine teacher quality on the basis of improved student achievement in the hope there would be less of an emphasis on defining teacher quality due to increased education, experience, or seniority in a district (Arens, 2005). These new systems equated quality with student outcomes (Barton, 2006). The expectation was to encourage states and districts to provide better professional development, placement, and recruitment (Arens, 2005).

Furthermore, the determination was the systems presented evaluations of schools or reforms that would translate into changes at the state level. The goal was to more effectively use available research (Arens, 2005; Barton, 2006). Arkansas requires districts to develop an improvement plan based on results of test data and create a plan of action. Each action statement in a school's Arkansas Consolidated School Improvement Plan (ACSIP) must be supported by research (ADE, 2005). Again, the logic behind the goal is that by providing specific data, schools would make decisions based on this information (Arens, 2005; Barton, 2006).

Lastly, the assumption was the provision of equitable opportunities for the benefit of all students (Arens, 2005). The intention was that data disaggregation will help to diagnose and treat deficiencies, which consequently upheld the primary principles of NCLB. The chief influence of accountability was that it spotlighted the attention of educators and policymakers on the need to adequately serve at-risk students (Goodwin, Englert, & Cicchinelli, 2003).

Models

The common theme in the previous studies specified that any rating system must incorporate data generated not just from one test but also from other measures of student and school success. In order to reduce criticism and negate unfairness found within the constraints of a one-shot test as the sole measure of student achievement, state policymakers should draw on several data sources (Popham, 2007b; Stapleman, 2000). State policymakers will use accountability results to determine if mandates are not only being enforced but succeeding as well (ADE, 2004). The accountability model unfortunately has become the primary information used to judge a district and subsequently punish or reward. Due to the potential sanctions, districts want results which clarify whether or not the teacher in a classroom is effective (Sokola, Weinberg, Andrzejewski, & Doorey, 2008). With all of the different stakeholders it is vital to know that the accountability system is reliable and the results offer accurate information.

Hopefully any accountability model implemented would have the consequences of higher test scores. However, the goal should be to produce as few unexpected consequences as possible and each model must be weighed accordingly. No model

will supply all of the desired outcomes or please all of the potential audiences. Each state must find a program to achieve academic goals, as well as providing accurate information to help with decision-making and encourage improved learning and higher student achievement (Goldschmidt & Choi, 2007). With unlimited examples of accountability models, states have the capability to use a cafeteria style of picking and choosing to best fit education needs.

In a study performed by the McREL, researchers polled public opinion on the definitions of different available models. A market model indicates that people have the right to vote with the ability to change school districts where students may leave low-performing schools (Arens, 2005). Performance models centered on a variety of assessment measures where goals were aligned and clearly stated. Regulatory models defer to fiscal accounting procedures and not on an accomplishment of standards (Arens, 2005).

There were two basic types of models that monitored school performance. The first was a status model which used a single year's assessment results as an indicator of school performance and passed rules based on these results (Goldschmidt & Choi, 2007). A more recent and now more popular model is the Growth Model. In a survey conducted by the American Association of School Administrators 40 percent of the superintendents believed schools should be allowed to use a growth model to measure achievement (Pascopella). While they wanted a true growth model that would drive teaching and learning, they were concerned over a lack of federal funds (Pascopella, 2008).

In November 2005, U.S. Secretary of Education, Margaret Spellings, announced a Growth Model pilot program where states submitted alternative accountability models to monitor schools (Hoff, 2008b). Arkansas recently reported they were one of seven states whose growth model plan was accepted by the United States Department of Education (DOE). The state will use this model in calculating AYP (ADE, 2007). DOE rejected several states that applied for the program because radical changes were initiated in the accountability plans. Several states' most significant proposal was to switch the order in which supplemental services and school choice are offered in schools failing to make AYP (Hoff, 2008b).

This change of philosophy by DOE displayed a willingness to recognize alternate variables affecting student achievement. Arkansas uses a version of the growth model in which schools have an increasing percentage of students scoring proficient on the state's Benchmark each year by the 2013-2014 school years, all students score proficient (ADE, 2007). According to the ADE (2007c), Questar Assessment Incorporated developed a model where students were matched by a strict set of conservative criteria using the spring 2006 Benchmark test administration as the base year. Questar matched students who had results on the spring 2007 Benchmark test administration and used the results in the growth model computation. For students with scores for only one year, the growth computations were completed, and the Growth Index and the Proficiency Threshold/Target values were stored for future application in determining student growth (ADE, 2007c).

Also, the model assessed the year-to-year growth of each child and determined which ones were making enough progress to achieve proficiency by eighth grade,

even though the performance level had not been rejected (ADE, 2007c). However, this method is more accurately defined as a path to proficiency model which does not enable a district to get credit for moving a special education eighth grade student from a third grade level to a fifth grade level even though this indicates something significant has taken place for the child.

Growth models use two or more years of assessment results as an indicator of school performance and make School Improvement decisions based on those results (Goldschmidt & Choi, 2007). DOE's Growth Model pilot program identified core principles that the program should target. Specifically, the growth models must set expectations for annual achievement based on meeting grade-level proficiency, which diminished the influence of student background or school characteristics (Hoff, 2009). Thus achievement targets for the 2013-2014 school year must be fixed at the same level for all students regardless of their characteristics or prior achievement levels. This places potentially unobtainable expectations for growth on initially poor performing students. A student may realize a growth of three grade levels in a year but still be below basic on a cut score proficiency test. No recognition is given to a child who is making remarkable gains even while he or she is still below grade level.

The idea of tracking individual student growth over time combined with the prohibition of aggregating estimated or observed growth for determining AYP makes it difficult to use these types of growth models. As a result, the two states adopting the pilot programs using the growth model created by DOE showed almost no change in the number of schools making or not making AYP compared with an existing status model (Choi, 2006). In a 2007 *Phi Delta Kappa/Gallup* poll, 82 percent of

Americans said they wanted schools rated on the improvement students make during the year, rather than the percentage who meet the state standard at the end of the year (Sokola, Weinberg, Andrzejewski, & Doorey, 2008).

While some states and researchers viewed the existing status and growth models as unsatisfactory, they advocated a value-added assessment to measure effective school and teachers (Choi, 2006; Doran & Fleischman, 2005). These models do not preclude measurement based on one set of test scores, but followed a student's improvement from one year to the next. Students test scores were converted statistically to a scale score so that achievement gains in one grade or one subject represented the same amount of growth in the same subject at the next grade level (Choi, 2006). Supporters believed it is important to account for the advantages and disadvantages students bring to the school because of prior instruction or a family situation (Toch, 2008).

There was some anxiety this method provided no diagnostic information for the teacher to use. Popham (2005b) stated, "Regretfully value-added methods sacrifice effective instructional diagnoses on the altar of statistical precision" (p. 84). However, the idea was to level the playing field using statistical methods. Another concern was these methods were too complex, and researchers recommended that before a state considers this method they consult professional, experienced research organizations (Doran & Fleishcman, 2005).

Accountability models are used as sanctions and rewards to drive reform. Examples of these rewards and sanctions are listed in the CRESST report, *Standards for Educational Accountability Systems*, with the advisement these carrots and sticks

started out broad and diffuse, then “move to specific consequences for individuals and institutions as the system aligns” (Baker, Linn, & Herman, 2002, p. 2). Stakeholders then have the opportunity to meet the requirements set by the models. However, there must be evidence of technical reliability with the measures used, and according to Englert, Fries, Martin-Glenn, and Michael (2005) error rates associated with misclassification of individuals or institutions should be published. They also discussed the requirement of an accountability model to align resources with the goals of the system. This alignment makes a difference in achievement especially with the more disadvantaged students. The redirection and assurance of equitable funding allows schools to focus on specific programs in need of improvement (Englert, et al., 2005).

Do accountability models work? The answer is both yes and no because it depends on the degree of rewards and sanctions built into the model. There has been evidence to support the models have pressured schools into change. David Figlio, in an August 19, 2008, on-line chat format through *Education Week*, cited examples of studies with which he was involved and the relationship to the success or failure of these models. He related a recent study with Cecillia Rouse, Jane Hannaway, and Dan Goldhaber in a working paper on the website of the National Center for the Analysis of Longitudinal Data in Educational Research where they found that schools subjected to greater accountability pressure tended to improve student test performance in reading and mathematics to a meaningful degree. Furthermore, research indicated that Florida schools responded to accountability pressures by changing some of their

instructional practices rather than inventing short-term test-taking tricks (Figlio, 2008).

Assessment Theory

Education reform over the last half century has been placed squarely on the shoulders of accountability and assessment (Crone, 2004; DOE, 2001). While testing and assessment have both critics and proponents, there were several reasons for the appeal of assessment with all of the players; 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 has been fairly inexpensive compared to other measures (Miles, 2001; Sokola, et al., 2008). Expensive items in place of testing/assessment measures involved hiring more certified staff as well as aides or increasing instruction time and reducing class size, which was unlikely as resources have already been stretched thin (Miles, 2001). The implementation of programmatic changes required significant professional development costs. All other things being equal, assessment was cheap (Miles, 2001).

A second reason for the appeal of testing and assessment as a reform tool was that policymakers were able to mandate targets. The philosophical idea, rightly or wrongly, is that an objective target score is a fair gauge of whether reform would be successful within a district (Barton, 2006). It was more difficult for school leaders to require and longer lasting deeper changes inside a classroom. Furthermore, testing and assessment on the surface are quick fixes for reform. This made it popular with Congress because the requirements were visible within an elected official's term in office (Loup & Petrilli, 2005). This may explain why NCLB received bi-partisan

support, and also why reauthorization did not take place until after the 2008 election. Lastly, assessment was appealing because results were easily reported to parents, the public, and the press (Fuller, Wright, Gesicki, & Kang, 2007). Scores started low and after a few years of testing; scores rose by the very nature of students and teachers becoming familiar with the assessment policies, even if nothing else was taking place in the school. This happened regardless of whether any fundamental changes were taking place in the achievement the assessment was designed to measure (Popham, 2003).

Testing and assessment have a variety of designs and forms, so first and foremost, the choices of which type to use are paramount to the goals of the assessment (Popham, 2003; Popham, 2007). Unfortunately, one potential problem was that there were often conflicting goals between local and state educators. Policymakers concentrated on the lowest performing schools and met those needs first. If blanket policies were implemented statewide, higher-performing schools would be reluctant to move away from programs that were already effective even though student achievement may not be maximized (Sokola, et al., 2008). Furthermore, a state would not want to employ strict guidelines while an individual school would want to use these stricter guidelines to force changes within the district (Lewis, 2000). Some reasons which may be for local or political rather than those which are educationally sound.

Those responsible for mandating and overseeing assessment reforms should be well-versed as to what the tests actually accomplish. It is crucial to apply assessments in the manner for which they were designed, especially if they are to be part of a

legislated accountability system (Sokola, et al., 2008). It is essential that educators follow the instructions in the test manual. Using a test for less than its intended purpose causes 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, et al., 2008 p. 27).

There are four factors to consider when choosing an assessment (Laitsch, 2005). The first item is the test type (Laitsch, 2005). There have been two primary types which included an achievement test or an aptitude test. Achievement and aptitude tests, while similar, measure two different concepts. Achievement tests measure the specific content a student has, or has not learned, whereas aptitude tests attempt to predict a student’s future behavior or achievement (Ravitch, 2007). The second factor is to determine what the test is going to be used for; diagnostic, placement, formative evaluation, or summative evaluation (Laitsch, 2005). The third is to question what scoring reference would 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 (Laitsch, 2005)?

Fourth, not only is the type of assessment key, but the value of the assessment is equally critical. James Popham (2003), emeritus professor of education at UCLA, provided three gauges as to whether an assessment had value. He referred to this as being instructionally sensitive. Popham’s definition of instructionally sensitive meant the test determined the presence of instructional improvement. The first indicator was

the degree of difficulty of the content standards measured. The second meter was the description of the tests assessed content standards, and the third gauge was the reporting procedures used for group and individual student reports (Popham, 2003).

Assessment for Learning

There are 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, 2007b). Assessment at the local level helped decide referral and screening decisions and supports decision-making for classification issues (Chappuis & Chappuis, 2008). For example, an assessment for learning may resolve whether a student was eligible for special services. Dietel (2005) stated “The task of the psychometrician today is not necessarily to test the child or youngster, but to examine the data for the processes of teaching and learning to generate the necessary assessment data that will promote learning” (p. 4).

Also this view incorporated a measurement for instructional planning decisions which helped to clarify and specify how and where a student was taught, or to identify if a student had mastered a set of subskills needed to move on to more difficult curricular goal. These tests are used to help teachers and administrators plan educational programs (Popham, 2006; Popham, 2007b). According to Chappuis and Chappuis (2008), assessment for learning helped to answer three questions for students: Where am I going? Where am I now? How can I close the gap? Feedback is the key because with this type of assessment there is still time to take action and

create a plan for students to get to where they need to be (Popham, 2006; Popham, 2007b).

Assessment of Learning

Assessment of learning occurs when students demonstrate knowledge of a particular curricular area for progress monitoring or grading purposes (Lewis, 2005). It is evaluative in nature and used for accountability, rewards, and sanctions. These assessments support student progress decisions (Popham, 2007b). A concern of educators has been the assessment of learning mandated by NCLB would overshadow assessment for learning as teachers focused on covering materials necessary to achieve AYP (Popham, 2006).

There are two general categories of assessments educators have used. The first was an informal assessment which means the collection of data by anything other than a standardized test (Starkman, 2006). These make up the majority used by the classroom teacher such as portfolios, teacher observation, teacher-made tests, and computer-based testing. Evaluations of this nature impart more accurate diagnostic information since they are not bound by the same constraints as statewide tests (Rabinowitz, 2005).

Informal Assessment

Informal assessments are also made up of three sub-groups: formative, interim or progress, and summative assessment. There is a great deal of confusion about the roles of these types of assessment (Starkman, 2006). What then is the difference? 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 (Popham, 2007b). Such assessments help guide instruction so that teachers and students have a general idea of what learning outcomes have been achieved, and what further focus is needed. It involves frequent testing, and a measurement of student learning is just one component (Chappuis & Chappuis, 2008.)

In 2004, the historic article “Inside the Black Box” written by Black and Wiliam in 1998 for *Phi Delta Kappan* which gave credence to formative assessment and its conclusions was revisited. The more recent article, “Working Inside the Black Box: Assessment for Learning in the Classroom,” written by Black, Harrison, Lee, Marshall, and Wiliam (2004) discussed the three questions which originated in the primary study. The first question asked if there was evidence that improving formative assessment raised standards, and the answer was still yes. The second question asked if there was room for improvement, and again, the response was yes. The last question asked if there was evidence about how to improve formative assessment (Harrison, Lee, Marshall, and Wiliam, 2004). This was where the two articles deviated. The updated article discussed that while new ideas emerged, there was enough detail that would enable teachers to implement these ideas in the classroom (Black, Harrison, Lee, Marshall, & Wiliam).

The second of the informal assessment subgroups is comprised of a more recent assessment term now known by the phrase interim assessment. The interim assessment is administered periodically throughout the year to monitor student progress (Perie, et al., 2007) toward meeting state standards, usually in math and literacy. These tests provide rapid, regular feedback to students, teachers, and

administrators. One indicator of the importance of interim/progress tests has been the rapid increase in availability of such products from commercial test providers (Marsh, Pane, & Hamilton, 2006). Approximately one hundred fifty districts throughout Arkansas use an interim assessment tool called the Target Test (O.U.R. Cooperative, n.d.). Students are evaluated periodically with a standards-based assessment and the results are provided within a few days. Ideally teachers would have immediate access to results and use them to drive instruction. If or how this is actually being done, would need to be studied and the success of the program remains to be seen. The district used in this study opted for an alternate interim assessment instead of the Target Test, but the premise is the same (O.U.R. Cooperative, n.d.).

Just like the teachers in Arkansas who may be using Target Test results to drive instruction, other formative tests would be equally ineffective should nothing happen after the assessment was complete (Popham, 2007b). Thirty-five years ago, Benjamin Bloom stressed the value of providing the student with feedback and the need to follow up with correctives (Guskey, 2008). He also stressed that these correctives must be fundamentally different from original instruction (Guskey, 2008). Lastly, in the informal sub-group is the summative assessment, which evaluates achievement at the end of specific educational programs (Popham, 2006; Ravitch, 2007). The purpose of the summative assessment is to measure the level of student, school, or program success. However, one problem has been that results were often reported in ways that made it difficult for teachers to comprehend, so even if the tests are suggested to use for formative purposes, a lack of teacher comprehension makes this difficult (Chappuis & Chappuis, 2008).

Formal Assessment

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 (Laittsch, 2005). 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). One fact that is hard to dispute is that testing is big business (Miles, 2001). This is an unregulated industry whose revenues have been skyrocketing. Not only is there a cost in the test itself, but the scoring and reporting of the tests are expensive (Cech, 2008b; Miles, 2001). Cech (2008a) reported, “Tests, test delivery, scoring, scoring analysis, professional development ... accounted for about 30 percent of the \$2.1 billion in overall assessment revenue” (p. 15).

Since the results of high-stakes test are so important, there has been a call to regulate them (Clarke, Madaus, Horn, Ramos, Lynch, & Lynch, 2001). Testing company executives reported that states spent \$700 to \$750 million annually on testing contracts (Toch, 2006). However, this equated to about one percent of the overall budget. As a result, tests have not been scrutinized as closely by the states and local districts as they should be. Many states do not have the time, finances, or staff to implement tests that align with state standards (Toch, 2006). These unaligned tests will give skewed results and lack validity (Toch, 2006). Unfortunately teachers have

been trying to cover mountains of standards they assume are on the test, but in the end the tests have covered completely different material (Kohn, 2001b).

As long as the federal government mandates testing and applies the funding carrot, states have no choice but to struggle daily to comply. In order to validate limited varying resources (time, money, staff,) local districts must employ these tests and disaggregate data to improve curriculum and instructional practices (Miles, 2001). Testing is only beneficial if the information gathered has transformed into practices that improve student learning. It is clear that “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 is difficult to determine a standardized assessment’s ability to enhance student learning, but even so, the quality of the assessment is paramount. It became even more problematic when states adopted the ideology that “test-based accountability systems embody the belief that public education can be improved through a simple strategy” (Stecher & Hamilton, 2002, p. 1). If states and local districts have been spending valuable time and money, but not yielding accurate information, precious resources were wasted (Herman & Baker, 2005). For as many different standardized tests available to the consumer, the more varied their ability to assess student knowledge.

There has been good news that often standardized tests undergo rigorous validation criteria, reliability testing, and standardization procedures from the testing company (Stecher & Hamilton, 2002). The rationale underlying reliability has been

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. In other words, chance effects should not have a significant influence on test scores (Runyon, Coleman & Pittenger, 2000). While reliability refers to whether test scores are constant indicators of student performance, validity signifies the degree to which the test items reflected the specified content domain (Ravitch, 2007).

There has been a concern that these large-scale external assessments will be unable to measure the academic content and curriculum covered at the local level (Wong & Nicotera, 2007). Furthermore, tests have drawn criticism from educators and policymakers who believed tests should not be used to make high-stakes decisions because they are limited in the ability to measure student attainment of high-quality academic standards (Wong & Nicotera, 2007).

. Educators must be familiar with the way each type of assessment operates in order to determine the multiple indicators of student performance. There must be enough information about instructional practices to make improvements (Wong & Nicotera).

In Arkansas, Questar Testing Company possessed the contract during the first two years of the study. Each item was field tested and then reviewed for bias. Items were developed that incorporated examples specific to Arkansas standards (Gray, 2007). The constructed response items included rubrics for scoring. The scorers originated in Arkansas. They were trained and graded blind in order to conceal and protect the names of the tested students. Each item had more than one scorer (Gray, 2007). Unfortunately, Arkansas, like most states, contracts with the testing company and when the contractual time runs out the testing company potentially changes. Harcourt

Pearson won the new contract, and will be the test manufacturer of the Arkansas Benchmark Test through 2013 (Gray, 2007).

The Association of American Publishers (AAP, 2000) believed standardized tests provided four critically important tasks: The first task is to identify the instructional requirements of individual students so educators 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 measure educational growth over time. Third, standardized tests should help to evaluate the effectiveness of educational programs. The fourth task is to monitor schools for educational accountability under NCLB. However, the AAP cautioned that tests should be considered a means to an end and not the end itself (AAP).

Even within the same category of standardized tests not all components have been equal. There are different question types and degrees of difficulty on individual tests (Laitsch, 2005). One common item format is multiple-choice. This type provides an adequate measure for lower level skills such as vocabulary and general principles (Laitsch, 2005). Constructed response offer the best gauge for complex achievement, such as application, inference, and generating hypotheses or conducting experiments (Laitsch, 2005).

Performance and portfolio assessments are not thought to be part of the standardized testing genre, but allow for a demonstration of student competency (Cavanaugh, 2008; Laitsch, 2005). In Arkansas, students with special needs are permitted to submit a portfolio to show proficiency in math and literacy when it is determined the regular test is not appropriate (ADE, 2008a). These include

performance assessments which offer presentations of student work. The portfolios 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, 2005). Individual states work with test companies to determine a design suitable for students with special needs.

There are two primary types of standardized tests: criterion-referenced tests and norm-referenced tests (Bond, 1996). Under NCLB, states may include either, or both, of these assessments, and the law also stated that 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 (DOE, 2003). 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, 2003).

Criterion-referenced tests are defined as student knowledge measured against a set of pre-determined standards (Ravitch, 2007). Educators choose these tests when they want to determine how well students master a set of skills or a desired curriculum. 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 (Ravitch, 2007). In Arkansas, the criterion-referenced assessment have been required by state statute, rule, or regulation and designed by the State to measure student performance/achievement on the State's Academic Content Standards (ADE, 2004).

Cut scores on 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. The test may be positively or negatively skewed depending on how well the teacher addresses the state-mandated content standards (Deubel, 2008). Critics would say this supports the argument for teaching to the test rather than teaching for student achievement (Deubel, 2008; Laitsch, 2005).

Norm-referenced tests are defined as student knowledge measured against other students in the 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, norm-referenced tests 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 (ADE, 2004) 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. The scores are, however, considered in the growth model to identify weaknesses based on score reports from the testing company (ADE, 2007c).

On a norm-referenced test, scores are reported so that half of the testers score in the top 50 percent and half in the bottom 50 percent (Laitsch, 2005). 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, score interpretations 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, 2005). In Arkansas the norm-referenced test previously given to students was the Iowa Test of Basic Skills (ITBS), this national test compared Arkansas and district students to the same subset nationwide. There were also problems with the first-year administration of this test. Harcourt accidentally sold the kindergarten test as a practice tests prior to the spring administration. The entire state's kindergarten scores were thrown out and these students were retested in the fall with the Metropolitan Achievement Test as first-grade students (D. Wolfe, August 05, 2008).

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 (2005) reported that the Association for Supervision and Curriculum Development (ASCD) advocated multiple measures as a gauge for the success of an accountability system. Laitsch suggested that ASCD supported assessments that are “fair, balanced, and grounded in the art and science of learning and teaching” (p. 3) and must be “reflective of curricular and developmental goals and representative of content those students have had an opportunity to learn” (p. 3). ASCD also focused on Limited English Proficient students and special needs students, and wanted an evaluation that would accommodate needs. Lastly, the assessment system should be “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” (Laitsch, 2005, p. 2).

Pre-Assessments

In a desire to predetermine student proficiency and achievement levels, schools have been creating or purchasing assessment systems to monitor student progress and determine how accurately students meet state standards throughout the year (Clarke, et al., 2001). In many states, reporting of annual scores are delivered too late in the year to accurately remediate student weaknesses, so the pre-assessments are essential to raise achievement levels. There are few assessments systems where the only purpose of the system is to predict performance on a later assessment. However, interest in these assessments will increase as the annual NCLB targets continue to rise (Perie, Marion, Gong, & Wurtzel, 2007). While the state tests have been important communicators of student achievement and allow schools to reform curriculum and instruction long-term, the tests do not provide ongoing information that schools may employ to incrementally improve instructional programs. Furthermore, the state tests do not address the learning problems of students with the most need (Herman & Baker, 2005). State tests are assessments of learning and districts should understand assessments for learning are a necessity to increase student achievement.

Carol Ann Tomlinson (2008) referred to informative assessments which guide instruction. Tomlinson stated:

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 allowed for educators to evaluate how students are performing at a single point in time, but if the results are reported immediately, and if the pre-assessments are administered at different points throughout the year, growth progress is measurable (Olson, 2007). 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 targeted instructional opportunities, it is necessary to align assessments to state-mandated content standards (Carter, 2007; Olson, 2007). This will in turn allow for growth measurement regardless of achievement status.

This new wealth of immediate student data presented educators with decision-making information. It permitted them to consider program decisions and evaluate teacher effectiveness. Olson (2007) explained:

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. (p 11)

According to 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 have been learning to use pre-

assessments and end-of-the-year test results to evaluate lack of, or increases, in student achievement (Reeves, 2004). This is a key change because most data-driven decision making a few years ago was more about looking at end-of-year test results with little or no analysis to tie-in causes. Pascopella (2006) explained, “It was an autopsy. I’ve never seen a patient get better because of an autopsy” (p. 40). A 2006 *Rand* study revealed a common set of factors to help explain why some educators tended to use data more and with greater levels of sophistication than others. The factors included; accessibility, quality (real or perceived), motivation, timeliness, staff capacity and support, and curriculum-pacing pressures (Marsh, et al., 2006).

Using data-driven decision-making does not guarantee effective decision-making (Englert, Fries, Martin-Glenn, & Michael, 2005). The process of translating data into information, knowledge, decisions, and actions is labor intensive, and practitioners need to consider the trade-off of time spent collecting and analyzing data, as well as the costs of providing needed support and infrastructure to facilitate data use (Marsh, Pane, & Hamilton, 2006).

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

The quality of the assessment has become essential, and “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 the pre-assessment benchmarks. The criteria, according to Herman and Baker, 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

In Memphis, students at Elmore Park Middle School use Think Link Inc.’s Predictive Assessment Series. The thirty-five minute tests closely mirror the content tested on Tennessee’s state-mandated benchmark tests and immediately rate a child’s performance. Students take the Think Link tests three times each year, and reports show the tactic is working. On a state report card Elmore Park raised its grade for value-added in math from an “F” to a “B” in two years, and raised its value-added grade in reading from “C” to “A” in one year (Sausner, 2005).

One available pre-assessment is the STAR Math test. Since program costs are considerable, it is essential that a district weigh the cost effectiveness against desired student achievement outcomes. This assessment is developed by the Renaissance Learning Company and offers computer-adaptive tests which provide the respondent with grade equivalency and percentile math scores for grades first through twelfth in less than fifteen minutes (Renaissance Learning, 2008). The accompanying Accelerated Math Program supports curriculum by providing individualized practice tailored to each student’s weakness, scoring responses automatically and delivering continuous feedback. All of the math questions are linked to recommendations

provided by the National Council on Teaching Mathematics (NCTM.) This is an important factor as the Arkansas student learning expectations are also closely aligned to the NCTM (Renaissance Learning, 2008).

The brochure published by the Renaissance Learning Company stated, “Teachers can predict achievement on state standards, determine the appropriate level of challenge, instantly place new students, identify those who need individual help, and plan individualized instruction on the skills students need to become successful at math” (Renaissance Learning, 2008, p. 3). A study which determined whether or not STAR Math predicted student achievement on the Arkansas Benchmark Test would allow a district to make informed financial decisions. Furthermore, if a range of scores were determined which had a high correlation of success on the Benchmark the information would be invaluable.

Consequences of State-Mandated Testing

In the current climate of mandated testing, it has been difficult to have a civil discussion about NCLB as proponents and dissenters weigh in. Reeves (2004), a centrist on testing issues who heads the Center for Performance Assessment based in Denver, discussed the myths associated with this legislation. He argued against the premise that this law was a Republican Party tactic to support vouchers and charter schools. His evidence was the Executive Order, signed by then President Bill Clinton, allowing parents to move their children out of schools failing to achieve adequate progress. In opposition to this view, Bracey (2004) argued, “The goal of NCLB is the destruction of public schools, not their salvation. NCLB sets schools up to fail and be privatized” (p. 68).

Positive Consequences

It may be impossible to find an educator 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 agreement across the board 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). Linn (2005), an education professor emeritus at the University of Colorado at Boulder and a frequent critic of NCLB, reviewed results of the legislation and 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. 5). Linn’s comments centered on a study of NCLB he participated in for the Center on Education Policy.

The language surrounding 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 make every effort to report performance in terms that are clear and understandable to students, parents, and the public (ADE, 2004; 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, 2000).

Those who support state-mandated standardized tests value these tests as tools in providing data and results necessary for schools to reform (Arens, 2005; Schmoker, 2000). 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). State tests are also powerful motivators for reform. Schools now have to set goals and evaluate their systems (Herman & Baker, 2005). This focus on accountability has led teachers to rethink their methodologies for teaching. The new concentration on standards included processes such as reasoning, problems solving, using multiple representations, communication, and making connections, which are embedded in math questions on standardized tests (Duebel, 2008).

Another positive result of testing has been the ability to focus on individual sub-groups and identify particular needs, since mandates also require these populations to meet AYP (Cavanaugh, 2008). Guilfoyle (2006) explained:

If nothing else NCLB has launched an unprecedented focus on the reading and math abilities of previously marginalized students. By requiring the disaggregation 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. (p. 11)

The accountability system attempts to assure adequate attention to these groups of students by requiring the separate reporting of results. Such disaggregated reporting of results provided a mechanism for monitoring the achievement of lower performing groups and narrowing achievement gaps (Linn, 2005). In Arkansas, the data have been disaggregated by sub-group. The state recognized, where the federal government

is just now waking up to this reality, that growth is as important as meeting AYP. If a sub-group showed ten percent growth from one year to the next, the school received safe-harbor to demonstrate that progress was made. As a result, the district would not be penalized (ADE, 2007a).

Negative Consequences

For every proponent of standardized testing there has been an equally vocal dissenter. Kohn, (2001a) the loudest of the critics, 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” (§ 1). State trends show there has been a positive statistical correlation between higher geographical areas socio-economic level and the level of proficiency ratings (NORMES, 2009).

Kohn (2004) also argued there have been no positive effects of testing. He believed 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, § 1). Other less emotional dissenters argued the test’s limitations, such as, the multiple-choice format which does not indicate a student’s ability to analyze in writing or apply processes (Schmoker, 2000).

Cech (2008b) quoted Koretz, a professor at Harvard’s graduate school of education, saying that due to the NCLB law, there has been widespread teaching to the test, strategic reallocation of teaching talent, and other means of gaming the high-stakes testing system. This produced scores on state standardized tests that were

substantially better than the students' mastery of the material. Arkansas has tried to overcome these limitations by providing questions which require written responses and mathematical open response questions forcing students to apply and infer (ADE, n.d.).

Another critic, Popham (2003), 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 (Wallace, 2000). This fear may drive schools to lose creativity and spend time teaching the techniques of test-taking rather than developing a more rigorous curricula (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. Principals face the possibility of losing 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 (Wallace, 2000).

There are opponents of NCLB who see the school choice legislation as being one step closer to 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).

NCLB presumed that by monitoring the percentage of students who are proficient in reading and mathematics, this would be sufficient to identify schools that are doing

a good job versus schools needing improvement (Nowak & Fuller, 2003).

Unfortunately, this assumption has several flaws. First, because schools are held accountable for performance by student subgroups, large diverse schools would be 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 scored at or above the proficient level in comparison with an annual target percentage places too much emphasis on student enrollment characteristics, such as 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 (Hoff, 2009). 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 vs. 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. 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 (Runyon, et al., 2000). The measurement error on standardized tests 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, 2004).

Another concern is that of test validity. Measurement experts have been explicit about what makes a test valid in an accountability system (Cech, 2008b; Popham, 2008). If alignment to the curriculum has been weak and instruction does not match the standards, then the assessment would not meet the standards for validity and the reported scores could not be relied on as an adequate judge of a school's effectiveness (Popham). 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 refers 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 (Cech, 2008b). If confidence intervals are used to calculate AYP where an error range is determined, either plus or minus, it will skew the results. The statistical measure of using a confidence interval would be correctly applied to sampling of a population and not on the complete population. The inaccurate use would provide an error range for an entire population who has already taken the test and is statistically inappropriate (Trochim, 2008). However, the federal government allows states to use this measure as way to keep fewer schools in the needing improvement phase (Popham, 2005a).

Less complex methods of loopholes to elevate AYP have been used. Often cut scores appear arbitrary, when states change them after raw scores have been reported,

or when the rigor of a test is weakened by making items easier (Popham, 2005a). Furthermore, schools tutor bubble-students, or those who fall just below the proficiency level, by teaching test taking techniques to move lower students upward. This practice does nothing to increase student comprehension of the standards. In extreme cases, low-performing students have been discouraged from attending on test day (Guilfoyle, 2006).

In Arkansas, the state mandated that 95 percent of any student population, combined or subgroup, must take the test, or the district or school will not meet AYP (ADE, 2004). When the eleventh grade literacy test was updated and rigor was increased, very few schools met AYP (ADE, 2004). As a result, the state revamped and lowered AYP target percents considerably. Otherwise, a majority of Arkansas high schools would be on the school improvement list (ADE, 2004).

Unintended Consequences

If accountability systems have the power to change behavior, as the early evidence indicated, then it is imperative to ensure that these systems change behavior in correct ways (Stecher & Hamilton, 2002). Occasionally high-stakes tests produce undesirable and unintended consequences, such as teaching the test or excluding students from testing (Fuhrman, 1999). Positive consequences of high-stakes testing include: better information about individual student's knowledge and skills, 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 (Cawelti, 2006). Negative consequences of high-stakes testing include: test frustration and discouragement, misplaced competitiveness causing students to devalue grades and

school assessments, and tying assessments to students' graduation or promotion whereby students drop out or increase the number of years necessary to graduate (Cawelti, 2006).

Positive consequences for teachers may include: a more efficient way to diagnose individual student needs and help teachers to identify areas of strength and weakness in the curriculum (Cawelti, 2006; Popham, 2007b). Furthermore, high-stakes testing may help identify content not mastered by students and redirect instruction. 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, 2006; Cawelti, 2006). Negative consequences for teachers may include encouraging teachers to focus on specific test content more than curriculum standards. 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 the students with practice problems as the teachers prepared the students for standardized testing (Cawelti, 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 (Popham, 2007a).

Positive consequences for administrators include: an examination of school policies related to curriculum and instruction, help administrators' judge program quality, lead them to change school policies to improve curriculum or instruction, and help them make better resource allocation decisions (Cawelti, 2006). Negative

consequences include: lead administrators to enact policies to increase test scores but not necessarily increase learning, cause administrators to reallocate resources to tested content at the expense of other courses, waste resources on test preparation (Stecher & Hamilton, 2002).

Accountability models will also have unintended consequences. Schools, in general, must be careful to overcome a hazardous application of concentrating on the bubble kids. This practice happens all too frequently and has become a negative, unintended consequence of testing. Neal and Whitmore, (as cited in Figlio, 2008) from the University of Chicago, who noted that accountability systems based on getting students above a given performance threshold tended to induce schools to focus on the kids who are almost proficient. Figlio (2008) stated:

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. (¶ 4)

Summary

The review of the literature indicated many researchers believe states need to develop accountability systems designed around several inputs of data rather than a one-shot test. There are a variety of growth models available for implementation which draws from a number of data sources. Furthermore, the review provided an in-depth examination of assessment theory. Assessment for learning is formative assessment theory which presents the educator with information about student understanding, enabling interventions to happen instantly (Popham, 2007b).

Assessment of learning is the summative evaluation system typically found in the state-mandated benchmarks (Popham, 2007b). The review presented consequences to mandated testing. In chapter three the methodology used to study the STAR Math test as a predictor of achievement was a quasi-experimental design. Data was collected and presented in chapter four. An analysis of the data and its impending implications for assessment were discussed in chapter five.

CHAPTER THREE – DESIGN AND METHODOLOGY

Introduction

A primary piece of No Child Left Behind (NCLB) was requiring states to test student populations. As a result, it is essential to be able to predict how students will perform on benchmark tests. This study was designed to discover if the STAR Math Test is an accurate indicator of student proficiency on the Arkansas Benchmark Test.

Several factors presented a rationale for this study. First, the STAR Math Test, in conjunction with its partner program, Accelerated Math, enables a teacher to offer remediation based on student weaknesses (Renaissance Learning, 2006). Second, state funding and control of a school district is based directly on benchmark performance (ADE, 2004). The third factor is fiscal responsibility where any purchased program must be determined to be worth the cost. If the program does not provide an accurate indicator then limited resources are wasted (Miles, 2001).

Population

Demographics for the school and district encompassing the three years accessed for the study have been shown on Table 1. All secondary data was available from a Northwest Arkansas Middle School, and was normally accessible to the researcher.

Table 1.

Demographics: Study Site School

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

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

Sub-populations < 10 students were not reported

Sampling Procedure

The data originated from the sixth, seventh, and eighth grade student populations and was compiled over a three year period. All student scores were kept anonymous for the purpose of the research. A random sampling was not appropriate for this study in order to limit interference from the nuisance variables. Runyan, et al. (2000) defined these as “Variables that may interfere with the assessment of the effects of the independent variable” (p. 17). In this study nuisance variable were those associated with students moving into the districts and include; outside curriculums and teachers.

In order to investigate how Arkansas educators viewed the use of pre-assessments as an indicator of student achievement, surveys were sent out via e-mail. These addresses were obtained through an administrators' list serve.

For the correlation between the 2005 STAR Math pre-test and the spring 2006 Benchmark, student scores from the spring 2005 Benchmark Test, the fall 2005 STAR Math pre-test, the spring 2006 Benchmark Test, and the spring 2006 STAR Math post-test. The sample size for the sixth grade included 82 students, and the seventh, and eighth grade sample sizes were 86 and 82 students, respectively. For the correlation between the 2006 STAR Math pre-test and the spring 2007 Benchmark, student samples participated in the spring 2006 Benchmark Test, the fall 2006 STAR Math pre-test, the spring 2007 Benchmark Test, and the spring 2007 STAR Math post-test. The sample size for the sixth grade included 97 students; the seventh had 69 students, and the eighth grades had 97 students. For the correlation between the 2007 STAR Math pre-test and the spring 2008 Benchmark, students participated in the spring 2007 Benchmark Test, the fall 2007 STAR Math pre-test, the spring 2008 Benchmark Test, and the spring 2008 STAR Math post-test. The sample size for the sixth grade included 117 students, the seventh at 91 students, and eighth grade had 95 students.

By including only these students in the study, new students moving into the district influenced by outside instructors and curricula did not affect the results. All students examined had an equal chance of being chosen based on the stated criteria. No students were included or excluded based on a sub-population status; chances of inclusion were exactly the same as the general population.

Research Setting

A consistent method used over the three years helped to limit nuisance variables. The STAR Math test was administered by a certified teacher in a computer-based laboratory. No outside help was available to the student, and the teacher acted only as a proctor for the testing session. Each tested group was given 45 minutes to complete the assessment, and students remained quiet until all others finished the test. The Arkansas Benchmark is a spring standardized test and was administered in the appropriate setting with certified staff, time constraints, and standardized procedures set by the state were strictly adhered to. Training for the Arkansas Benchmark was provided by the same District Test Coordinator over the three-year time span (Conner, 2009)

Research Design

Questions

Four questions were addressed in this study

1. What relationship exists between the STAR Math pre-test and post-test in the sixth, seventh, and eighth grades for 2006, 2007, and 2008?
2. What relationship exists between corresponding student populations over two consecutive years of the Arkansas Benchmark in the sixth, seventh, and eighth grades for 2006, 2007, and 2008?
3. What relationship exists between the STAR Math pre-test and the Arkansas Benchmark examination in the sixth, seventh, and eighth grades from 2006, 2007, and 2008?

4. How do Arkansas administrators view the use of pre-assessments as an indicator of achievement on the Arkansas Benchmark Test?

Independent Variable

The independent variable in the study was the Star Math pre-test scores.

Dependent Variable

The dependent variable in the study was the Arkansas Benchmark Test scores of corresponding students.

Hypothesis

Null hypothesis

There is no significant relationship between the STAR Math Test and student achievement on the Arkansas Benchmark Test. The null hypothesis is designated by the symbol H_0

Alternate hypothesis

There is a significant relationship between the STAR Math Test and student achievement on the Arkansas Benchmark Test. The alternate hypothesis is designated by the symbol H_1

Timeline

STAR Math test data and Arkansas Benchmark scores for three years spanning 2005 to 2008 were gathered in the fall of 2008. At the same time, surveys were sent across Arkansas via e-mail. After data collection, the information was analyzed and presented in the spring of 2009.

Table 2.

Timeline of the Study

<u>Date</u>	<u>Event</u>
Spring 2005	Middle school students participate in the Arkansas Benchmark
Fall 2005	STAR Math pre-test given to all middle school students
Spring 2006	Middle school students participate in the Arkansas Benchmark
Spring 2006	STAR Math post-test given to all middle school students
Fall 2006	STAR Math pre-test given to all middle school students
Spring 2007	Middle school students participate in the Arkansas Benchmark
Spring 2007	STAR Math post-test given to all middle school students
Fall 2007	STAR Math pre-test given to all middle school students
Spring 2008	Middle school students participate in the Arkansas Benchmark
Spring 2008	STAR Math post-test given to all middle school students
Fall 2008	Surveys sent to Arkansas Educators
Fall 2008	Data are gathered and analyzed
	Statistical models included frequency distributions, correlations, tests for reliability, Coefficients of Determination, and scatter plots

The STAR Math computerized program scored both pre-tests and post-tests by running scan sheets through a scantron machine linked to a computer by software. Results were compiled and available through a local web-based program supported by a password. Prior to 2008, the Questar testing company scored the Benchmark Tests and returned results to the district by individual, school, and district reports. Scorers throughout Arkansas were trained using scoring guides and rubrics. In 2008, Harcourt Pearson developed, implemented, and scored the new augmented Benchmark Tests also using trained scorers from within the state. In both cases, all open-response questions were scored blind. Results are returned to individual districts by May 31st of each year (ADE, 2007; Gray, 2007).

The first procedure was to separately test the reliability of each variable; the STAR Math Test and the Arkansas Benchmark examination. This enabled the researcher to determine the extent to which each variable individually produced a consistent outcome from year to year. For the STAR Math Test, correlations were calculated with the aid of the SPSS Statistical Software Program using pre-tests and post-tests from the available student sample populations. This analysis measured pre-tests and post-tests spanning three years using ordinal ranks for each of the sixth, seventh, and eighth grades. Calculations were done separately for all three years. The same procedure was repeated to test the reliability of the Arkansas Benchmark Test. The spring 2005 Benchmark and the spring 2006 Benchmark were ranked from the existing student populations and correlated. This procedure was repeated for the spring 2006 and the spring 2007 Benchmarks as well as the spring 2007 and spring 2008 Benchmarks.

The procedure examined the relationship between the STAR Math test and the Benchmark students by ranking samples according to their grade equivalency on the STAR Math pre-test and according to their corresponding spring 2006 Benchmark assessment scores. Students' scale scores on the benchmark were converted to raw score percents prior to ordinal ranking. 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. Scores included each of the sixth, seventh, and eighth grade levels.

Strategies Applied in the Study

The research design implemented for this study was a quasi-experimental design receiving this designation because a random sampling assignment was not applied (Trochim, 2008). The design also incorporated multiple groups and multiple waves of measurement in order to ensure a triangulation of data. Two types of triangulation were used in the study. The first was data triangulation which involved space, time and persons (Triangulation in Educational Research, n.d.). This study used data over three separate years and three separate grade levels with each unit measured independently. The second was methodological triangulation, which involved using more than one method and consisted of within-method or between-method strategies (Triangulation in Educational Research, n.d.). The following methods were used in the study.

Coefficient of Determination

This is a technique used to interpret the correlation coefficient designated by the symbol r^2 , and is defined as the percentage of variance in one variable that can be

described or explained by the other variable (Runyon, et al., 2000). The Coefficient of Determination figured the effect the independent variable, the STAR Math pre-test, had on the dependent variable, the spring Benchmark Assessment. The standard was set where the alpha-level (α) = $r^2 > 40\%$ and was considered necessary to reject the H_0 and accept the H_1 .

Descriptive Statistics

A set of statistical procedures used to organize, summarize and present data (Runyon, et al., 2000).

Frequency histogram. One of the descriptive statistics implemented in the study. It is a form of a bar graph representing a frequency distribution in which a continuous line or bars indicates the frequency of each score or group of scores (Runyon, et al., 2000). For this study, the strategy was applied to research the apparent ceiling effect evident when using finite scores and ordinal ranks.

Mean. A measure of central tendency calculated by adding all of the scores in a data set and dividing by the number of scores (Runyon, et al., 2000). In this study, the mean provided a basis for comparison between the grade levels and the years studied. It was applied on the raw data sets and the ordinal ranking data sets.

Standard deviation. A descriptive statistic used for reporting where approximately two-thirds of the distribution lies (Runyon, et al., 2000). It was calculated by finding the square root of the variance. A normal, unskewed curve will have 34 percent of the cases between the mean and 1 standard deviation above or below the mean; 68 percent of cases between 1 standard deviation above and 1 below the mean; 95.5

percent of cases will be within two standard deviations of the mean (Medical University of South Carolina, n.d.).

Linear Regression

This strategy was used when the projections were expected to be in a straight line with actual values (Griffith, 2007). In this study, *curve estimation* was applied. The curve can be used to estimate the values of points not yet in the data set. Specifically this was done through *extrapolation* which is defined as extending the curve beyond the existing points (Griffith, 2007).

Omega squared

Omega squared was an index of the degree to which the variance in one variable accounts for the variance in another (Runyon, et al., 2000). Omega squared was calculated by using the following formula.

$$w^2 = \frac{df_{\text{between}}(F - 1)}{df_{\text{between}}(F - 1) + N}$$

The standard was set where $a = w^2 > 40\%$ and was considered necessary to reject the H_0 and accept the H_1 .

One way ANOVA

This is a form of an analysis of variance that allowed the researcher to compare the effects of different levels of a single variable. “The purpose of the ANOVA test is to determine the existence (or nonexistence) of a statistically difference *among* the group means” (Brase & Brase, 2006, p. 722). The results were reported in table format and analyzed through the significance level (Runyon, et al., 2000). The standard was set where $a = p < .05$ and was necessary to reject the H_0 .

Pearson Correlation Coefficient

This was the primary statistical model used in the study and this statistic allowed the researcher to describe the extent to which the data fit a linear model. The coefficient ranged in value from -1 to +1. Zero indicates no relationship between the independent and dependent variable from Gay and Airasian's book *Educational Research: Competencies for Analysis and Applications* (as cited in Wisdom, 2008). The closer the coefficient is to the value of one, the closer the variable values are to fitting a perfectly straight line when graphed on the x-y coordinate plane from Hinton's work "Statistics Explained" (as cited in Wisdom, 2008). The primary data sets were ordinal rankings and as such a *Spearman Correlation Coefficient* is the statistical tool normally used.

However, when there are numerous tied ranks on either, or both the X- and Y-variables, the Spearman formula tends to yield spuriously high coefficient of correlation. When there are many ties, it is preferable to apply the Pearson *r* formula to the *ranked* data. (Runyon, et al., 2000, p. 188)

A tied rank refers to the fact that the number of the sample size was larger than the available ranks, so that within the ordinal ranked group several had the same rank score. Since there were numerous ties with both the X- and Y-variables, the primary correlation dedicated for this study was a Pearson correlation but using ranked instead of raw data. Ties were averaged and the mean was calculated for the ranked dependent variables. However, for comparison purposes the study included correlations for Spearman using ranked data and Pearson using raw data. All data was run through the SPSS Graduate Pack software to reduce potential calculation errors.

The standard was set where $a = r > .500$ and was considered necessary to reject the H_0 and accept the H_1 .

Survey

Surveys were collected from around the state and the results were compiled to garner further information from Arkansas educators. Questions were created by the researcher and designed to evaluate educators' views of the ability of a pre-assessment to predict student achievement on the Arkansas Benchmark examination. This survey was based on a stratified sampling. This is a technique in which the entire population is divided into distinct subgroups or strata, based on specific characteristics (Brase & Brase, 2006). In this case all members of the sample group had at least a bachelor's degree and experience in public education.

Statistical Treatment of Data

Magnitude of the Relationship

There are two basic features of every relation between variables. These are the relations of magnitude (size) and reliability (truthfulness) (Elementary Concepts in Statistics, n.d.). The magnitude of the independent variable over the dependent variable is uncovered through correlation calculations "where an attempt to somehow evaluate the observed relation by comparing to the maximum imaginable relation between specific variables (Elementary Concepts in Statistics, n.d., ¶ 3). The independent variable in this study was the Star Math test. The dependent variable was the student results on the Arkansas Benchmark Test.

As a note, correlation research does not try to influence any variables, but only measures them to look for relations between the set of variables (Brase & Brase, 2006). Data from correlation research can only be interpreted in causal terms based on theory, but cannot conclusively prove causality (Runyon, et al., 2000). A Line of Best Fit was graphed to determine if there was a linear relationship between the ordinal ranks of the independent and dependent variables (Griffith, 2007).

Reliability of the Relationship

Reliability or truthfulness of the hypothesis is calculated by determining the statistical significance or *p-value* of the variables over time (Trochim, 2008). The statistical significance of a result uncovers the degree to which the result is true. However, a research finding may be true without being important (Creative Research Systems, 2007-2009). The higher the *p-value*, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population (Runyon, et al., 2000).

Alpha-level

The alpha-level (α) represents the level of significance set by the experimenter. It is the confidence with which the researcher can decide to reject the null hypothesis (Runyon, et al., 2000). The significance level is the probability value used to conclude that the null hypothesis is an incorrect statement. In this study the statistical significance was calculated separately for each of the three years using multiple measures. The *p-value*, the measured probability of a finding occurring by chance alone given that the null hypothesis is actually true, is set at <0.05 which converts to a 95% confidence interval of how likely the sample mean represents the population

mean (Medical University of South Carolina, n.d.). The alpha-level was set at the following standards by the researcher $r > .500$, $p\text{-value} < .05$, $r^2 > 40\%$ and $w^2 > 40\%$. All four standards must be met to conclusively reject the H_0 and accept the H_1 and complete the methodological triangulation of data (Triangulation in Educational Research, n.d.).

Ethical and Political Considerations of the Study

While all information was available to the appropriate district personnel, all student names remain confidential. As a result, no personal student information appears in this study. In addition, survey respondents were kept confidential.

Summary

Great care was taken with the design and methodology of this study. A quasi-experimental design was chosen to fit the standards of data and methodological triangulation. Three years worth of data was accumulated for three separate grade levels. Multiple measures were also involved in the research design. While Pearson with ordinal rankings was the primary correlation, additional measures were applied to be used as a comparison. Other strategies included; descriptive statistics, linear regression, omega squared, and analysis of the variance and a survey. Furthermore, separate tests of internal reliability were performed separately on both the Benchmark and the STAR Math tests to ensure the variables were reliable by themselves. In addition, nuisance variables were considered and limited to the best of the researcher's ability.

CHAPTER FOUR – RESULTS

Introduction

There were several factors to consider before the results of the study were interpreted. The first was the sample selection for the study. The samples were taken from the sixth, seventh, and eighth grades for the years 2005-2006, 2006-2007, and 2007-2008. Only samples which included students who had taken the previous year's benchmark test, the STAR Math pre-test, the STAR Math post-test, and the current spring benchmark test for the observed years were considered. As a result the sample size was different for each year and each grade level. To complete a data triangulation each grade and each year was measured independently and then compared rather than calculated as a whole. This allowed the range of the results as well as the reliability to be considered.

The second factor was the trustworthiness of the study. As with any study this was limited by the nuisance or extraneous variables. These were examined in chapter one under the limitations of the study. Primarily, there were important outside characteristics to student achievement such as teacher quality, curriculum quality, parental involvement, socio-economic status, and language barriers. It was impossible to control these variables within the constraints of this study, but every effort was made to minimize their effect.

Another factor which might affect the results included the dependent variable, the benchmark test. For example, the degree of difficulty was changed from year to year,

and the cut scores identifying proficiency were also adjusted periodically (ADE, 2007a). To overcome the latter problem and increase the trustworthiness of the study, proficiency ratings were disregarded and scale scores were converted to raw scores and percentages were calculated. Furthermore, at the end of each contract period the test manufacturer potentially changes (ADE, n.d.), indicating the necessity for a measurement to ensure the new test and the previous benchmark test still had a reasonable degree of reliability.

Results

Three years worth of data were accumulated, and the magnitude and reliability of the variables were calculated by using a Pearson r correlation, a scatter plot for line of best fit and curve estimation, a coefficient of determination, an analysis of the variance (ANOVA) test, and a calculation of Omega squared. Based on the application and analysis, the original null hypothesis was deemed to be incorrect. The H_0 was rejected when all of the calculated correlation coefficients were above the .500 mark and the p -values calculated were below the .01 to .05 level of significance. The results of the statistical calculations were consistent when comparing each separate grade level and each of the three testing cycles. Therefore, the H_1 was accepted that a statistical significance does exist between the independent and dependent variables.

Analysis of Data

Research Question Number One

What relationship exists between the STAR Math pre-test and post-test in the sixth, seventh, and eighth grades for 2006, 2007, and 2008?

Internal Reliability of the Independent Variable

The first step was to determine the reliability of the independent variable. In order to calculate the relationship between the pre-tests and post-tests of the independent variable, it was necessary to measure the correlation of the STAR Math pre-test to the post-test from 2005-2006, 2006-2007 and 2007-2008 in the sixth, seventh, and eighth grades. Tests for reliability using the Pearson r correlations were performed to measure for consistent outcomes. Table 3 showed a statistically significant correlation of reliability in more than one result with a range of .625 to .836. The sixth grade showed the least significant change from one pre-post test year to the next with consistent declines for all three years. The eighth grade pre-test and post-test for the 2007-2008 year was the only grade in the analysis which experienced an increase in the correlation over the previous year. Sample sizes were greater than the number indicated because the grade equivalencies were reduced to ordinal rankings which only allowed for ranks between 1 and 12.9 causing numerous ties among the samples creating the appearance sample sizes were reduced.

Table 3.

STAR Math Test Correlations of the pre and post tests

Grade	STAR Math Pre-2005 Post-2006 Tests Correlation Coefficients	STAR Math Pre-2006 Post-2007 Tests Correlation Coefficients	STAR Math Pre-2007 Post-2008 Tests Correlation Coefficients
Sixth	.719*** n=49	.707*** n=50	.625*** n=62
Seventh	.838*** n=51	.717*** n=43	.673*** n=47
Eighth	.806*** n=48	.639*** n=40	.707*** n=45

Note: Pearson using ordinal ranks – Sig. (2 tailed)

n = sample size

*** Correlation is significant at the .500 level

Research Question Number Two

What relationship exists between corresponding student populations over two consecutive years of the Arkansas Benchmark in the sixth, seventh, and eighth grades for 2006, 2007, and 2008?

Internal Reliability of the Dependent Variable

The second step in the analysis was to discover the reliability of the Arkansas Benchmark from 2005-2006, 2006-2007 and 2007-2008 in the sixth, seventh, and eighth grades. The 2005 spring administration of the Benchmark was compared to the

2006 spring administration of the Benchmark using corresponding student populations. The results in Table 4 showed statistically significant correlation coefficient calculations when testing reliability in more than one measure of the Arkansas Benchmark Test with a range of .794 to .938. The results were greater than were found in the STAR Math pre-test and post-test correlations. However, there was a decline in each progressive grade level from the spring 2007 to the spring 2008 administration of the Benchmark.

The change in the test manufacturer for the spring 2008 administrations of the Benchmark test which might account for the decline, but even with the decrease the correlation coefficients demonstrate a statistically significant result over time. This indicated minimal effect from the limitation of introducing a different test manufacturer as was discussed in chapter one. Furthermore, this also illustrated the necessity of performing a measurement to ensure that the dependent variable provided for consistent outcomes on its own merit.

Table 4.

Arkansas Benchmark Correlations

Grade	Spring 2005 and 2006 Benchmark Coefficients	Spring 2006 and 2007 Benchmark Coefficients	Spring 2007 and 2008 Benchmark Coefficients
Sixth	.938*** n=45	.894*** n=48	.794*** n=60
Seventh	.960*** n=38	.917*** n=41	.859*** n=43
Eighth	.900*** n=40	.904*** n=46	.871*** n=49

Note: Pearson using ordinal ranks – Sig. (2 tailed)

n = sample size

*** Correlation is significant at the .500 level

Research Question Number Three

What relationship exists between the STAR Math pre-test and the Arkansas Benchmark examination in the sixth, seventh, and eighth grades from 2006, 2007, and 2008?

Descriptive Statistics of the Raw Data

The initial treatment of the variables was to examine the raw data of the STAR Math grade equivalency and the raw score percents of the Benchmark for each of the three years and three grade levels. Comparisons factored in the following: sample group sizes; mean grade equivalencies; standard deviations for the STAR Math pre-

test; frequency distributions of the grade equivalencies; mean raw score percents for the Benchmark Test; the standard deviation of the benchmark test; and frequency distributions of the raw score percents.. When the standard deviations were compared between the grade equivalencies and the benchmarks, the benchmark standard deviations were significantly larger. The higher the standard deviation, the more different scores were from one another and from the mean (Runyon, et al., 2000). These facts were summarized in Tables 5, 6, and 7. Furthermore when the frequency distributions have a high standard of deviation, the mean is not a good measure of central tendency (Runyon, et al., 2000). Over the three year period, the range of the means between the grade equivalencies and the average raw score percents went from a low of 6.14 and a high of 65.22. The same held true for the range of the means of the standard deviations with a low of 2.04 and a high of 19.052.

Since the grade equivalencies and raw score percents available on the scale were finite, frequency distributions were graphed to provide a visual display of the actual spread of the data. This clearly showed any areas where ceiling or floor effects played a role in data analysis. When comparing the frequency distributions of the grade equivalencies the normal curve distribution was disrupted due to the ceiling of a 12.9 grade equivalency. However, the average raw score percents presented a relatively normal curve distribution. The Frequency Distributions of the raw data were displayed in Appendix A, and Figures A1 through A18. The ceiling effect is first evident in Figure A5.

Table 5.

Comparison of Sample Sizes, Means and Standard Deviations of STAR Math Tests and Benchmark Tests

Grade	Sample Group Size	<u>Year</u>		<u>2005-2006</u>	
		Mean GE STAR Math pre-test	SD STAR Math pre-test	Mean Raw Score Percent Benchmark test	SD Benchmark test
Sixth	82	6.14	2.04	59.10	17.214
Seventh	86	6.98	2.23	47.08	16.504
Eighth	82	8.24	2.72	47.72	15.182

Note: GE = Grade Equivalency

Table 6.

Comparison of Sample Sizes, Means and Standard Deviations of STAR Math Tests and Benchmark Tests

Grade	Sample Group Size	<u>Year</u>		<u>2006-2007</u>	
		Mean GE STAR Math pre-test	SD STAR Math pre-test	Mean Raw Score Percent Benchmark test	SD Benchmark test
Sixth	97	7.30	2.894	65.22	18.58
Seventh	69	8.17	3.232	54.44	16.961
Eighth	97	8.77	3.092	46.21	16.105

Note: GE = Grade Equivalency

Table 7.

Comparison of Sample Sizes, Means and Standard Deviations of STAR Math Tests and Benchmark Tests

Grade	Sample Group Size	<u>Year</u>		<u>2007-2008</u>	
		Mean GE STAR Math pre-test	SD STAR Math pre-test	Mean Raw Score Percent Benchmark test	SD Benchmark test
Sixth	117	6.28	2.477	63.33	17.583
Seventh	91	8.66	3.294	50.88	16.646
Eighth	95	7.47	3.233	45.63	19.052

Note: GE = Grade Equivalency

Descriptive Statistics of the Ordinal Data

Once the means, the standard deviations, and the frequency distributions were calculated and evaluated, and the raw data of the samples were compared, it was essential to repeat the process for the ordinal rankings of the sample population. This was necessary when it was apparent the spread of the standard deviations and means of the raw data were too wide. The comparison of the raw data compared apples (STAR Math grade equivalencies) to oranges (Benchmark raw score percents). The

comparison of the ordinal rankings placed the data as a comparison of apples (Ordinal ranks of grade equivalencies) to apples (Ordinal ranks of benchmark raw score percents) which is displayed in Tables 8, 9, and 10. This summary showed a similar average of the means and a similar average of the standard deviations between the variables. Over the three year period, the range of the means between the ordinal ranks of the grade equivalencies and the ordinal ranks of the average raw score percents went from a low of 21.33 and a high of 26.84. The same held true for the range of the means of the standard deviations with a low of 9.723 and a high of 14.954. Based on the results of the descriptive statistics, it was established the ordinal ranks rather than the raw data would produce more reliable results. The Frequency Distributions showed the ceiling effects were present when the levels are finite. This is also evident in the ordinal rankings of the raw score percents, however it did not appear in the previous frequency distributions of the raw score percent prior to the ordinal rankings. Figures B1 through B18 in Appendix B summarize the ordinal ranked information.

Table 8.

Comparison of Sample Sizes and Means of Ordinal Ranks of STAR Math Tests and Ordinal Ranks of Benchmark Tests

Grade	Sample Group Size	<u>Year</u>		<u>2005-2006</u>	
		Mean Ordinal Ranks	SD Ordinal Ranks	Mean Ordinal Ranks	SD Ordinal Ranks
		STAR Math pre-test	STAR Math pre-test	Benchmark test	Benchmark test
Sixth	82	26.84	12.723	24.51	11.984
Seventh	86	25.50	13.173	21.33	9.723
Eighth	82	26.83	13.938	22.07	10.452

Table 9.

Comparison of Sample Sizes and Means of Ordinal Ranks of STAR Math Tests and Ordinal Ranks of Benchmark Tests

Grade	Sample Group Size	<u>Year</u>		<u>2006-2007</u>	
		Mean Ordinal Ranks	SD Ordinal Ranks	Mean Ordinal Ranks	SD Ordinal Ranks
		STAR Math pre-test	STAR Math pre-test	Benchmark test	Benchmark test
Sixth	97	25.67	14.118	21.26	13.001
Seventh	69	18.67	13.757	21.33	10.849
Eighth	97	15.66	12.872	25.78	12.072

Table 10.

Comparison of Sample Sizes and Means of Ordinal Ranks of STAR Math Tests and Ordinal Ranks of Benchmark Tests

Grade	Sample Group Size	<u>Year</u>		<u>2007-2008</u>	
		Mean Ordinal Ranks	SD Ordinal Ranks	Mean Ordinal Ranks	SD Ordinal Ranks
		STAR Math pre-test	STAR Math pre-test	Benchmark test	Benchmark test
Sixth	117	19.04	13.864	25.06	12.15
Seventh	91	19.69	14.954	23.60	11.825
Eighth	95	18.67	13.445	25.06	12.15

Linear Regression Results

Once the descriptive statistics and frequency descriptions were completed, the next step was to complete a line of best fit to determine whether or not linear regression models were the proper choice as a statistical test. Furthermore, with a collection of data points, it is possible to create a curve that passes through or very near those points. The curve can be used to estimate the values of points not yet calculated

(Runyon, et al., 2000). The graphic presentation of values is not as numerically accurate as a table of numbers, but it has some advantages. “Predictions are only estimations no matter how sophisticated, so presenting a prediction as a graph is as good as with numbers even with the inherent inexactness” (Griffith, 2007, p. 240-41). The line of best fit and curve estimation was seen on a scatter plot of the raw data of the STAR Math pre-tests and spring Benchmark Tests. Each dot represents the relationship of the grade equivalencies on the STAR Math test measured to the raw score percents of the Arkansas Benchmark Test for the sixth, seventh, and eighth grades in the 2005-2006, 2006-2007, and 2007-2008 school years’.

The predicted values are represented in three ways. The linear interpretation is the best fit of a straight line to the dots. The line that passes closest to each of the points is called the regression line. The quadratic line is the best fit of a line that curves in one direction. The cubic line reverses the direction of its curve in an attempt to fit as closely as possible. None of the curves fit the data points exactly, but they give the best possible prediction of the result (Runyon, et al., 2000). The Figures 1 through 9 displayed the information necessary to determine that the data does exhibit a linear pattern.

Y-axis Dependent Variable 2006 Benchmark Raw Score Percents - Sixth Grade

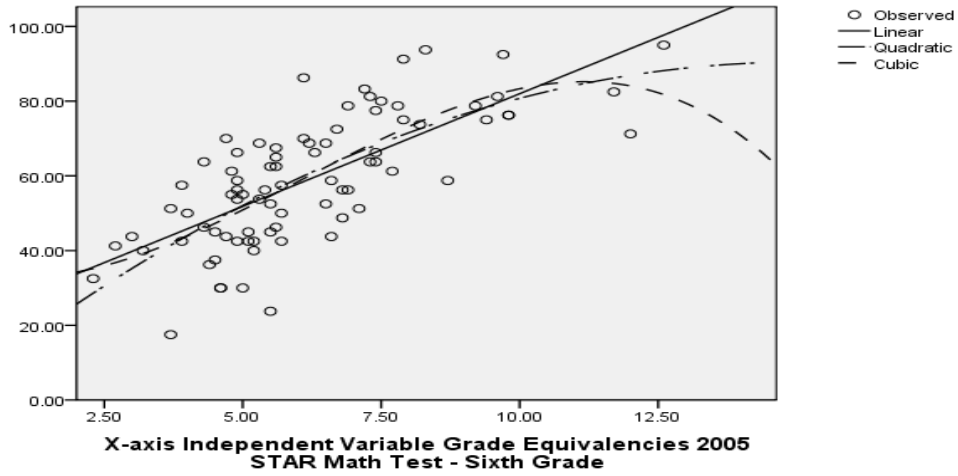


Figure 1. Curve Estimation for the Independent Variable and the Dependent Variable for the Sixth Grade 2005-2006

Y-axis Dependent Variable 2006 Benchmark Raw Score Percents - Seventh Grade

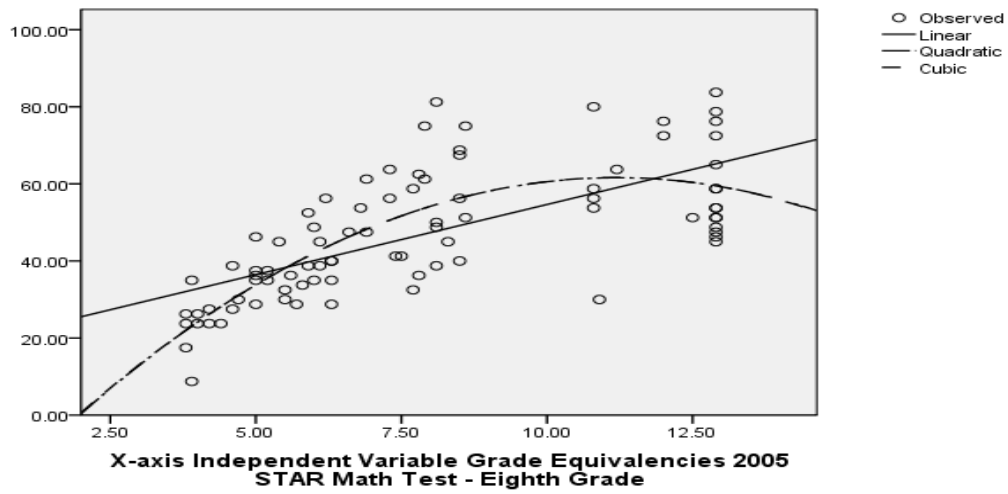


Figure 2. Curve Estimation for the Independent Variable and the Dependent Variable for the Seventh Grade 2005-2006

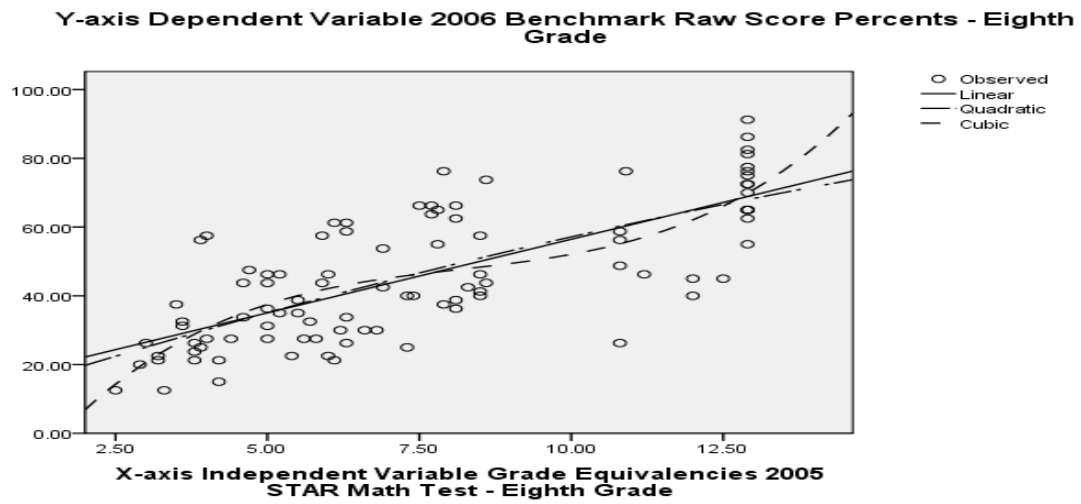


Figure 3. Curve Estimation for the Independent Variable and the Dependent Variable for the Eighth Grade 2005-2006

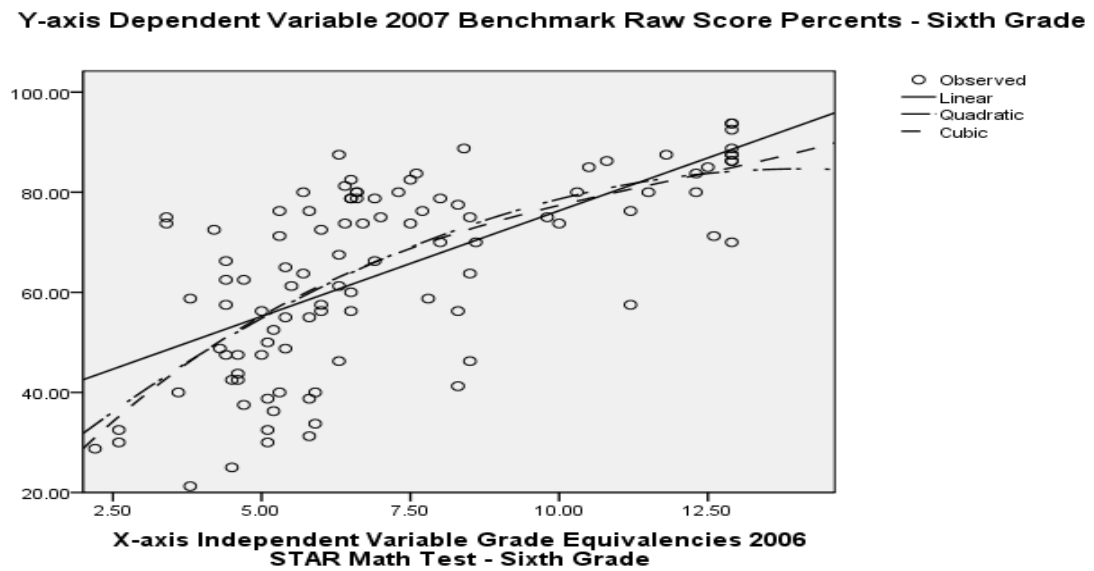


Figure 4. Curve Estimation for the Independent Variable and the Dependent Variable for the Sixth Grade 2006-2007

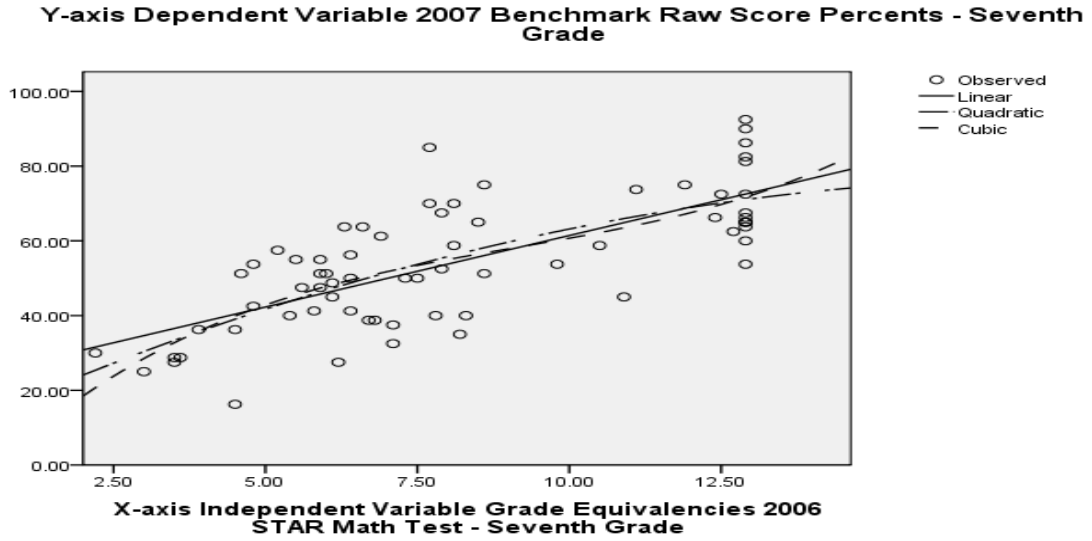


Figure 5. Curve Estimation for the Independent Variable and the Dependent Variable for the Seventh Grade 2006-2007

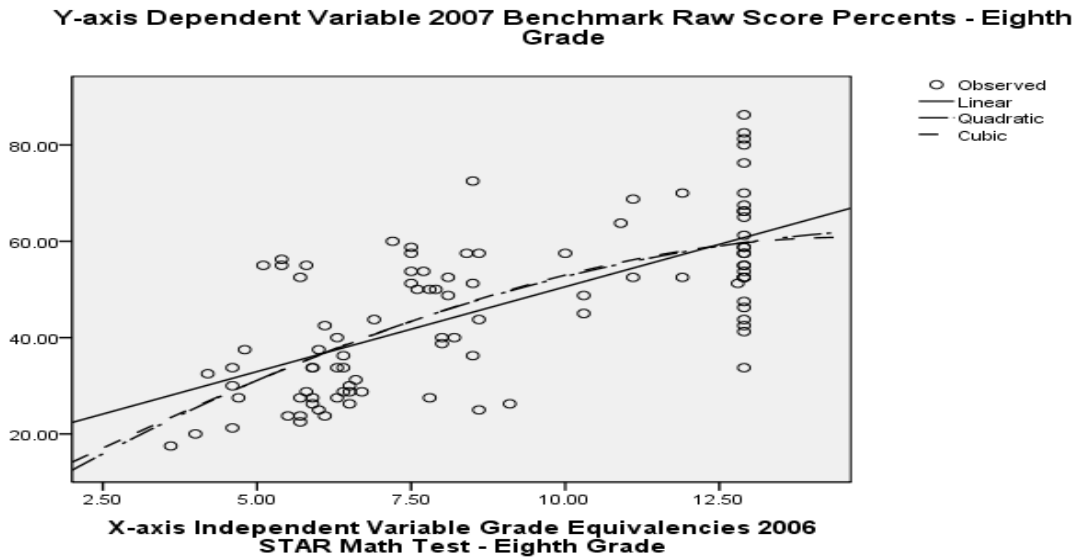


Figure 6. Curve Estimation for the Independent Variable and the Dependent Variable for the Eighth Grade 2006-2007

Y-axis Dependent Variable 2008 Benchmark Raw Score Percents - Sixth Grade

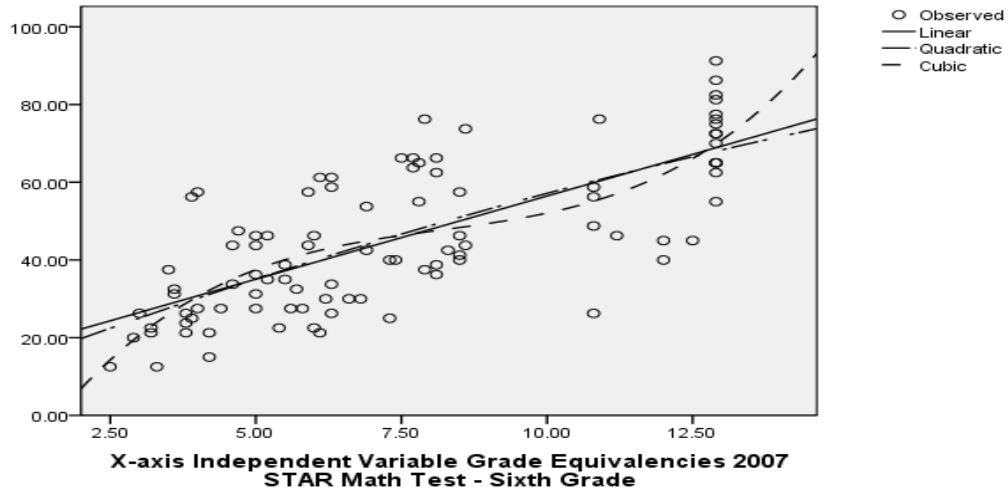


Figure 7. Curve Estimation for the Independent Variable and the Dependent Variable for the Sixth Grade 2007-2008

Y-axis Dependent Variable 2008 Benchmark Raw Score Percents - Seventh Grade

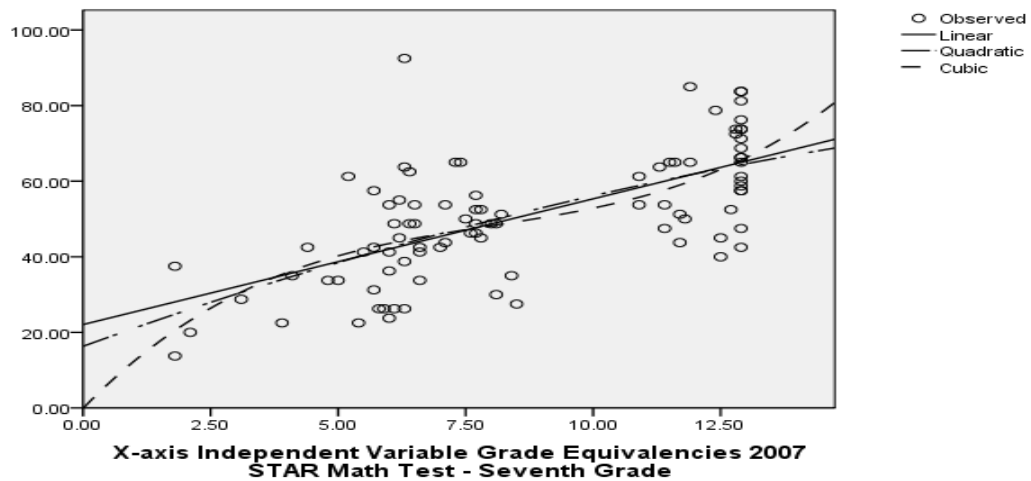


Figure 8. Curve Estimation for the Independent Variable and the Dependent Variable for the Seventh Grade 2007-2008

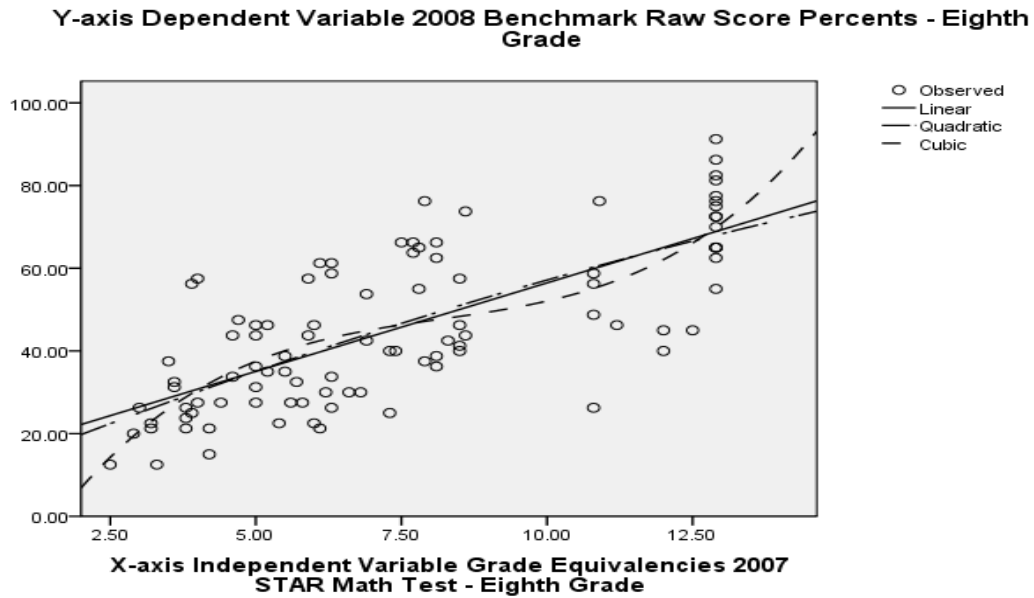


Figure 9. Curve Estimation for the Independent Variable and the Dependent Variable for the Eighth Grade 2007-2008

Correlation Analyses

Once it was determined that the raw data of the independent and dependent variables presented in a linear model, the primary test of the relationship a Pearson r correlation was deemed appropriate. This was accomplished by computing the correlation coefficient between the STAR Math pre-test and the spring Benchmark examination for the sixth, seventh, and eighth grades in the 2005-2006, the 2006-2007, and the 2007-2008 school years'. While correlation is a measure of direction and degree of relationship between two variables, a correlation coefficient is a numerical index of that relationship (Runyon, et al., 2000).

Calculations were completed for all three grade levels individually, and even though the primary correlation was Pearson using ordinal ranking due to the large number of ties among the raw data (Runyon, et al., 2000): correlations for Pearson using raw data and Spearman using ordinal rankings were also calculated as a reference. Tables 11, 12, and 13 displayed the results of all three correlation applications.

In Table 11 a statistically significant relationship, defined as not due to chance (Creative Research Systems, 2007-2009), existed between the ordinal ranks of grade equivalence on the STAR Math pre-test in the fall of 2005 to that of the ordinal ranks of the percentage obtained on the spring administration of the Arkansas Benchmark Test in the spring of 2006. The range of the primary correlation was .760 to .842. The sixth grade showed a slightly higher correlation between the ranking of grade equivalency and the Benchmark Test while the eighth grade presented the smallest of the correlations. When comparing the different correlation types, there was a small change between Pearson using ordinal ranks and Spearman. There was a more significant decrease when Pearson using raw percents was applied. However, all three applications showed a mid to high correlation between the STAR Math pre-test and the ordinal ranks of the percentage obtained on the spring administration of the Arkansas Benchmark Test.

Table 11.

Correlations for STAR Math Fall Pre-test and Spring Benchmark Test

Grade	<u>Year</u>		<u>2005-2006</u>
	Pearson using ordinal ranks – Sig. (2 tailed)	Spearman using ordinal ranks	Pearson using raw percents – Sig. (2-tailed)
Sixth	.842*** n=49	.846*** n=49	.720*** n=82
Seventh	.831*** n=51	.819*** n=51	.685*** n=86
Eighth	.760*** n=50	.746*** n=50	.750*** n=82

Note: n = sample size

*** Correlation Coefficient is significant at the .500 level

While slightly smaller correlations existed in Table 12 between the STAR Math pre-test in 2006 and the Benchmark in 2007, there was still a relatively high correlation between the ordinal ranks of grade equivalence on the STAR Math pre-test to that of the ordinal ranks of the percentage obtained on the spring administration of the Arkansas Benchmark Test. The sixth grade showed a slightly higher correlation between the ranking of grade equivalency and the Benchmark Test while the eighth grade presented the smallest of the correlations.

Table 12.

Correlations for STAR Math Fall Pre-test and Spring Benchmark Test

Grade	<u>Year</u>		<u>2006-2007</u>
	Pearson using ordinal ranks – Sig. (2 tailed)	Spearman using ordinal ranks	Pearson using raw percents – Sig. (2-tailed)
Sixth	.746*** n=49	.737*** n=49	.643*** n=97
Seventh	.665*** n=43	.641*** n=43	.729*** n=69
Eighth	.658*** n=40	.653*** n=40	.676*** n=97

Note: n = sample size

*** Correlation Coefficient is significant at the .500 level

In Table 13 there was a slightly higher correlation between the STAR Math pre-test in 2007 and the Benchmark in 2008, and there continued to be a relatively high correlation between the ordinal ranks of grade equivalence on the STAR Math pre-test to that of the ordinal ranks of the percentage obtained on the spring administration of the Arkansas Benchmark Test. The sixth grade showed a slightly higher correlation between the ranking of grade equivalency and the Benchmark Test while the seventh grade presented the smallest of the correlations. It is important to note, this is the testing year where the test manufacturer changed.

Table 13.

Correlations for STAR Math Fall Pre-test and Spring Benchmark Test

Grade	<u>Year</u>		<u>2007-2008</u>
	Pearson using ordinal ranks – Sig. (2 tailed)	Spearman using ordinal ranks	Pearson using raw percents – Sig. (2-tailed)
Sixth	.721*** n=45	.582*** n=45	.728*** n=117
Seventh	.661*** n=47	.663*** n=47	.658*** n=91
Eighth	.721*** n=45	.728*** n=45	.727*** n=95

Note: n = sample size

*** Correlation Coefficient is significant at the .500 level

Coefficient of Determination

Next, the coefficient of determination (r^2) was calculated and converted to a percent for the sixth, seventh, and eighth grades between the ordinal ranking of the independent variables and the ordinal ranking of the dependent variable. This was intended to further determine the strength of the correlation coefficient (Runyon, et al., 2000). The r^2 was calculated to establish the effect the STAR Math Test, the independent variable, had on the Arkansas Benchmark Test, the dependent variable. This information was summarized in Tables 14, 15, and 16. The results are mixed but the trend reflected the r^2 was reduced at each grade level and was also lower each of the years studied. The sixth grade all three years was the highest with the range from 71 percent down to 52 percent. The seventh grade moved from 69 percent to 44 percent over the next two years. The pattern changed with the eighth grade starting at 58 percent and dropping to 43 percent and moving back up to 52 percent in the 2007-2008 school year.

Table 14.

Coefficient of Determination between the STAR Math pre-Test and the Arkansas Benchmark Test

Grade	<u>Year</u>	r^2	<u>2005-2006</u>
	r		%
Sixth	.842***	.709	71%
Seventh	.830***	.689	69%
Eighth	.760***	.578	58%

Note: *** Correlation Coefficient is significant at the .500 level

Table 15.

Coefficient of Determination between the STAR Math pre-Test and the Arkansas Benchmark Test

Grade	<u>Year</u>	r^2	<u>2006-2007</u>
	r		%
Sixth	.746***	.557	56%
Seventh	.665***	.442	44%
Eighth	.658***	.434	43%

Note: *** Correlation Coefficient is significant at the .500 level

Table 16.

Coefficient of Determination between the STAR Math pre-Test and the Arkansas Benchmark Test

Grade	<u>Year</u>	r^2	<u>2006-2007</u>
	r		%
Sixth	.721***	.520	52%
Seventh	.661***	.437	44%
Eighth	.721***	.520	52%

Note: *** Correlation Coefficient is significant at the .500 level

Analysis of Variance

An additional method of determining the effect the independent variable has on the dependent variable was to perform an analysis of variance test or ANOVA. The purpose of this was to test the differences in means for statistical significance. This was accomplished by analyzing the variance, that is, by partitioning the total variance into the component that is due to true random error and the components that are due to differences in the means (Runyon, et al., 2000). These latter variance components were then tested for statistical significance and, if significant, the null hypothesis was rejected, and the alternative hypothesis was accepted (Brase & Brase, 2006).

In every calculation for Tables 17 through 25 the F_{observed} is greater than the F_{critical} and in each case the $p\text{-value}$ or statistical significance level was less than .01 which equated to a statistically significant determination (Runyon, et al., 2000). Furthermore only Table 18 fell in the above category. All other calculations were less than .005 which is considered highly significant. When the $p\text{-value}$ is less than .01, the H_0 is rejected (Elementary Concepts in Statistics, n.d.). “Specifically, the size of the F -ratio and $p\text{-value}$ indicate only whether we can reject the null hypothesis given the value selected for H_0 ” (Runyon, et al., 2000, p. 372). The H_1 was then accepted.

Table 17.

One way ANOVA test Sixth grade

	<u>Year</u>			<u>2005-2006</u>	
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9332.404	48	194.425	2.789	.001***
Within Groups	2300.083	33	69.699		
Total	11632.488	81			

*Note: *** Significant at the <.05 p-value*

Table 18.

One way ANOVA test Seventh grade

	<u>Year</u>			<u>2005-2006</u>	
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5934.234	48	123.630	2.178	.008***
Within Groups	2100.650	37	56.774		
Total	8034.884	85			

*Note: *** Significant at the <.05 p-value*

Table 19.

One way ANOVA test Eighth grade

	<u>Year</u>	<u>2005-2006</u>			
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10493.071	43	244.025	3.677	.000***
Within Groups	3384.550	51	66.364		
Total	13877.621	94			

*Note: *** Significant at the <.05 p-value*

Table 20.

One way ANOVA test Sixth grade

	<u>Year</u>			<u>2006-2007</u>	
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	13586.417	48	283.050	4.181	.000***
Within Groups	3655.506	54	67.695		
Total	17241.922	102			

*Note: *** Significant at the <.05 p-value*

Table 21.

One way ANOVA test Seventh grade

	<u>Year</u>			<u>2006-2007</u>	
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7135.905	42	169.902	5.093	.000***
Within Groups	867.429	26	33.363		
Total	8003.333	68			

*Note: *** Significant at the <.05 p-value*

Table 22.

One way ANOVA test Eighth grade

	<u>Year</u>			<u>2006-2007</u>	
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9800.213	39	251.288	3.418	.000***
Within Groups	4190.241	57	73.513		
Total	13990.454	96			

*Note: *** Significant at the <.05 p-value*

Table 23.

One way ANOVA test Sixth grade

	<u>Year</u>			<u>2007-2008</u>	
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10663.738	44	242.358	3.770	.000***
Within Groups	3213.883	50	64.278		
Total	13877.621	94			

*Note: *** Significant at the <.05 p-value*

Table 24.

One way ANOVA test Seventh grade

	<u>Year</u>			<u>2007-2008</u>	
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9483.647	46	206.166	2.926	.000***
Within Groups	3100.111	44	70.457		
Total	12583.758	90			

*Note: *** Significant at the <.05 p-value*

Table 25.

One way ANOVA test Eighth grade

	<u>Year</u>			<u>2007-2008</u>	
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10493.071	43	244.025	3.677	.000***
Within Groups	3384.550	51	66.364		
Total	13877.621	94			

*Note: *** Significant at the <.05 p-value*

Omega Squared

In order to evaluate the degree to which the independent variable is associated with the dependent variable, it is important to convert the *F*-ratio of the One way ANOVA to Omega squared (Runyon, et al., 2000). Table 26 displays the result of those calculations. Figures for these calculations were taken from Tables 16 through 24 which displayed the ANOVA treatments.

Table 26.

Omega squared results

Year	<u>2005-2006</u>	<u>2006-2007</u>	<u>2007-2008</u>
Grade			
Sixth	51%	60%	56%
Seventh	40%	71%	49%
Eighth	55%	49%	55%

Predictive Abilities

A final step in the analysis was to determine whether or not a range of STAR Math scores can supply performance indicators on the Arkansas Benchmark. For example in the sixth grade can a grade equivalency of 6.14 indicate a student would have a raw score percent of 50 or better? Initial calculations could not determine this range. The only thing which can be stated with a degree of accuracy was that the higher ordinal ranking of the grade equivalency the greater the ordinal ranking of raw scores percent.

The mathematics cut score in the sixth grade on the Arkansas Benchmark was 46 out of 80 or a 58 percent. Unfortunately the sample size over three years was too small with a total of fourteen scores available at the cut score percent, and the range was too wide from a grade equivalency of 3.3 to 8.7 at the matching 58 percent cut point. These figures included the 2005-2006, 2006-2007 and 2007-2008 school years. As a result, there was no identifiable range of sixth grade STAR Math pre-test scores which predetermine sixth grade benchmark scores.

The mathematics cut score on the Benchmark in the seventh grade was a 38 out of 80 or a 48 percent. Unfortunately the sample size over three years was too small with a total of five scores available at the cut score percent and the range too wide from a grade equivalency of 6.1 to 9.0 at the matching 48 percent cut point. These figures included the 2005-2006, 2006-2007 and 2007-2008 school years. As a result, there was no identifiable range of seventh grade STAR Math pre-test scores which predetermine seventh grade benchmark scores.

The mathematics cut score on the benchmark in the eighth grade was a 39 out of 80 or a 49 percent. Unfortunately the sample size over the three years was too small with a total of eight scores at the cut score percent and the range too wide from a grade equivalency of 7.0 to 10.8 at the matching 49 percent cut point. This includes the 2005-2006, 2006-2007 and 2007-2008 school years. As a result, there was no identifiable range of eighth grade STAR Math pre-test scores which predetermine eighth grade benchmark scores. A range of scores can be determined with a larger sample size for all three grades.

Research Question Number Four

How do Arkansas administrators view the use of pre-assessments as an indicator of achievement on the Arkansas Benchmark Test?

Survey

Ninety-two educators responded out of the approximately 200 surveys e-mailed to area educators. Over 40 schools were represented and only two respondents had duties in more than one building. Forty-eight work in an elementary setting. Six from the intermediate, fourteen from the middle school and twenty-eight from high school answered the survey. Twenty-six of the 38 or 68 percent who responded that their schools used the STAR Math Test also stated they believed pre-assessments accurately provide a predictor for student achievement. Thirty of the 50 respondents or 60 percent stated their school used a grade-level pre-assessment at the beginning of the year also believe pre-assessments accurately provide a predictor for student achievement on the Arkansas Benchmark Test.

Table 27.

Survey Results – Area Arkansas Educators

<u>Question</u>	<u># of Yes Responses</u>	<u># of No Responses</u>
Does your school use grade-level pre-assessments at the beginning of the year in math to determine student achievement levels?	50	42
Does your school use the STAR Math Test form the Renaissance Learning Company?	38	54
Do you believe pre-assessments accurately provide a predictor for student achievement on the Arkansas Benchmark Test? (Please answer whether or not your district uses a pre-assessment test.	58	34

Note: Total number of responses = 92

Deductive Conclusions

Based on the results, it can be stated that the higher the grade equivalency a student scored on the STAR Math pre-test, there will be a statistically significant correlation that the same student will score a comparative raw score percent on the spring administration of the Arkansas Benchmark Test. The H_0 stated the STAR Math Test is not an accurate predictor of student achievement on the Arkansas Benchmark Test, and with these conclusions the H_0 was rejected and H_1 stating, the STAR Math Test is an accurate predictor on the Arkansas Benchmark Test, was accepted. However, it is necessary to note the results did not provide a specific range of scores students could expect to attain on the Benchmark based on the STAR Math grade equivalency. This would be valuable data and worthy of future study.

Summary

The results of the study indicated a strong relationship between the ordinal rankings of the STAR Math pre-test and the ordinal rankings of the Benchmark test. There was a significant correlation between a high ordinal ranking for the pre-test and high ordinal ranking for the raw score percent on the Benchmark Test. “If two variables are known to be strongly related, we can predict one from the other” (Trochim, 2008, ¶ Analysis). However, the results did not explain a causal relationship. While there was a temporal directionality, meaning the independent variable occurred in time before the dependent variable (Runyon, et al., 2000), there was the issue of the third variable problem. This means there were too many events which could occur between the pre-test for STAR Math and the spring Benchmark Test (Runyon, et al., 2000). The study proved there was significant magnitude and

reliability between the two variables but did not prove causality. The results were also strengthened by the standards set in using data and methodological triangulation.

CHAPTER FIVE – DISCUSSION

Introduction

As stated in chapter one, the rationale for this study was to help Arkansas districts achieve AYP as state funding is directly linked to test scores. Districts accurately predicting student achievement can target student weakness prior to the benchmark test and focus efforts on direct remediation. If target areas are identifiable, districts can restructure the curricula more effectively and efficiently. An additional, but no less important reason for predicting student achievement is student placement in honor classes. In Arkansas, benchmark scores are required to be returned to districts by the last day of June. However, this is long after student schedules have been designed for the next school year. Accurate student placement is vital to ensuring student success in future courses, and improper placement will prove frustrating and increase opportunity for student failure.

The focus of the research was the STAR Math pre-assessment. However, since program costs are considerable, it is essential that a district weigh the cost effectiveness against desired student achievement outcomes. A study which determined whether or not STAR Math predicted student achievement on the Arkansas Benchmark test allows a district to make more informed financial decisions. Furthermore, if a range of scores could have been determined to have a high correlation of success on the Benchmark the information would be invaluable.

As discussed in chapter four, the first step was to test the internal reliability of each of the variables used in the study. Once it was determined that both the independent and dependent variables produced consistent outcomes over time the magnitude and reliability of the variables were calculated. The degree to which the independent variable - the STAR Math test had on the dependent variable – the Arkansas Benchmark Test was also measured. The magnitude was calculated through the Pearson r correlations using the ordinal ranks of the STAR Math grade equivalencies and the ordinal ranks of the Benchmark raw score percents. The Coefficient of Determination was also factored to determine the size.

Reliability was calculated with a one way ANOVA test to measure the statistical significance level of the variables. Next Omega squared measured the degree to which the variance in one variable accounted for the variance in another. Finally a graphic representation of the line of best fit and curve estimation was displayed.

A quasi-experimental design was chosen to allow for a stratified sample. By choosing this sample it enabled the study to limit the nuisance variables. Only students who had completed the previous years' benchmark, a STAR Math pre-test, a STAR Math post-test 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

By measuring both the magnitude and reliability of the variables a reasonable conclusion was that there is a positive significant correlation between the STAR Math

pre-test and the Arkansas Benchmark Test. This was important because it provided answers to questions identified in the rationale for the study. In essence, what can a set of test scores tell about the quality of education and the relationship to student performance? As a result of this question, the overarching problem was to find a statistically significant predictor of student achievement that can be monitored over time and used as a source for remediation and early intervention. Now that student weaknesses are identifiable, math curriculums can be restructured to be more efficient and cost-effective. Programs that do not serve students' best interests are not necessary, and student placement in various math classes can be aided with the use of STAR math. While the Renaissance Learning Company also produced a STAR Reader test, this study did not include this program, nor draw any conclusions concerning this application.

The format of authentic research is important for school districts to participate in. Research that is practical and provides answers to questions about specific programs or curricula currently in use or being considered by the district is invaluable. With the advent of No Child Left Behind public educators must use resources wisely, so that limited funds are spent in the most effective manner.

Recommendations

Due to the positive statistical significance of the results correlating the STAR Math test to the Arkansas Benchmark test, the researcher recommends continued use of the STAR math program within the district involved in the study. It is also advised that further research with additional data be completed. An investigation to uncover a line of regression, or the ability to predict Y - the raw score percent- based on a

distinct STAR Math grade equivalency would be invaluable. The researcher would further recommend that public educators participate in a study of any particular program or curriculum that is under consideration. This is the only way to ensure that goals are being met.

Summary

The results of this study indicated a strong relationship between the ordinal rankings of the STAR Math pre-test and the ordinal rankings of the Benchmark test. There is a positive correlation between a high ordinal ranking for the pre-test and high ordinal ranking for the raw score percent on the Benchmark Test. However, as stated in chapter four, these results do not explain a causal relationship. This study proves that there is significant magnitude and reliability between the two variables, but this in and of itself does not prove causality. However, it does suggest and support the continued use of the STAR Math test to predict student achievement on the Arkansas Benchmark Test.

REFERENCES

- Arens, S. (2005, July). *Examining the meaning of accountability: Reframing the construct*. (Issues Brief), Aurora, CO: Mid-continent Research for Education and Learning.
- Arkansas Department of Education. (n.d.). *Arkansas comprehensive testing, assessment, and accountability program*. Retrieved February 15, 2009, from <http://arkansased.org/testing/assessment.html>
- Arkansas Department of Education. (2004, June). *Rules governing the Arkansas comprehensive testing, assessment and accountability program*. ADE 188-1, Retrieved June 03, 2008 from <http://arkansased.org>
- Arkansas Department of Education, (2005, July). *Consolidated state application accountability plan*. ADE News release, Retrieved September 08, 2008 from <http://arkansased.org>
- Arkansas Department of Education. (2007a, July). *Rules governing the Arkansas comprehensive testing, assessment and accountability program*. ADE 188-1, Retrieved January 06, 2009 from <http://arkansased.org> Amended from June 2004.
- Arkansas Department of Education. (2007b, September). *Arkansas students performance holds steady on the "Nation's Report Card"*. ADE News release, Retrieved September 08, 2008 from <http://arkansased.org>
- Arkansas Department of Education. (2007c, October). *Arkansas growth model positively impact adequate yearly progress count*. ADE News release, Retrieved November 12, 2007 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, June). *Equating study final concordance tables between Stanford 10 and Iowa Test of Basic Skills for grades 1-9*. Prepared by Pearson for the ADE, Retrieved February 10, 2009 from http://arkansased.org/testing/pdf/sat10-itbs_concordance_063008.pdf
- 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.
- Baker, E.L., Herman, J.L., & Linn, R.L. (2005, Winter). *Evidence-based rationales for assessment systems*. (CRESST Line, 2,7). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Barton, P. (2006, November). Needed: Higher standards for accountability. *Education Leadership*, 64(3), 28-31.
- Black, P., & Wiliam, D. (1998, October). Inside the black box: Raising the standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-48.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, Retrieved October 10, 2008 from www.phideltakappan.org

- Bond, L. A. (1996). Norm- and criterion-referenced testing. *Practical Assessment, Research and Evaluation*, 5(2). Retrieved August 21, 2006 from <http://pareonline.net>
- Bracey, G. (2004). *Setting the record straight* (2nd ed.). Portsmouth, NH: Heinemann. (Original work published 1997).
- Brase, C., & Brase, C. (2006). *Understandable statistics: Concepts and methods*. Boston, MA: Houghton Mifflin.
- Carter, L. (2007). *Total instructional alignment: From standards to student success*. Bloomington, IN: Solution Tree.
- Cavanaugh, S. (2008, July). Testing officials tackle accommodations and exclusions for special student populations. *Education Week*, 27(43), 1,16.
- Cawelti, G. (2006, November). The side effects of NCLB. *Education Leadership*, 64(3), 64-68.
- Cech, S.J. (2008a, September). Test industry split over 'formative' assessment. *Education Week*, 28(4), p.15.
- Cech, S.J. (2008b, October). Testing expert sees 'illusions of progress' under NCLB. *Education Week*, 28(6), p. 8.
- Chappuis, S., & Chappuis, J. (2008, January). The best value in formative assessments. *Education Leadership*, 65(4), 14-18.
- Choi, K. (2006). *A new value-added model using longitudinal multiple-cohorts data*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

- 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>
- Conrad, J.K. (2008, July). *Wagging the dog: Leaning into the future of state assessment and accountability* (Trend of the Month). Aurora, CO: Mid-continent Research for Education and Learning.
- Creative Research Systems. (2007-2009). *Significance in Statistics and Surveys*. Retrieved February 19, 2009 from <http://www.surveysystem.com/signif.htm>
- 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>.
- Dietel (2005, Fall). *Testing to inform learning*. (CRESST Line, 4-6). Los Angeles, CA: University of California. National Center for Research on Evaluation, Standards, and Student Testing.
- Doran, H., & Fleischman, S., (2005, November). Challenges of a value-added assessment. *Education Leadership*, 63(3), 85-86.
- Deubel, P. (2008, April). Accountability, yes. Teaching to the test, no. *The Journal*, Retrieved October 7, 2008 from <http://www.thejournal.com>

- Elementary Concepts in Statistics. (n.d.). *How to measure the magnitude (strength) of relations between variables*. Retrieved February 19, 2009 from <http://www.statsoft.com/textbook/esc.html>
- Englert, K., Fries, D., Martin-Glenn, M., & Michael, S. (2005, November). How are educators using data: A comparative analysis of superintendent, principal, and teachers' perception of accountability systems. *Regional Educational Laboratory*, Contract #ED-01-CO-0006.
- Figlio, D. (2008) Testing and accountability in the NCLB era. *Education Week*, Retrieved August 26, 2008 from 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.
- Goldschmidt, P., & Choi, K. (2007, Winter). *The practical benefits of growth models for accountability and the limitations under NCLB*. (Policy Brief 9), Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Goodwin, B., Englert, K., & Cicchinelli, L. (2003, February). *Comprehensive accountability systems: A framework for evaluation*. (Revised edition). Aurora, CO: Mid-continent Research for Education and Learning.
- Gray, J. (2007, August). *ACTAAP updates*. Presentation conducted at the Arkansas Association of Educational Administrators in Little Rock, Arkansas.

- Griffith, A. (2008). *SPSS for dummies*. Hoboken, NJ: Wiley Publishing Company.
- Guskey, T. (2008, January). The rest of the story. *Education Leadership*, 65(4), 28-34.
- Guifolye, C. (2006, November). NCLB: Is there life beyond testing? *Education Leadership*, 64(3), 8-13.
- Herman, J.L., & Baker, E.L. (2005, November). Making benchmark testing work. *Educational Leadership*, 63(3), 48-54.
- Herman, J.L., & Choi, K. (2008, August). *Formative assessment and the improvement of middle school science learning: The role of teacher accuracy* (Crest Report 740). Los Angeles, CA: University of California. National Center for Research on Evaluation, Standards, and Student Testing.
- Hoff, D. J. (2008a, July). 2 new coalitions seek influence on campaigns. *Education Week*, 27(42), 1, 24.
- Hoff, D. J. (2008b, July). NCLB leeway allows states to hone plans. *Education Week*, 27(43), 1, 26.
- Hoff, D. J. (2008c, October). NCLB debate at the sidelines. *Education Week*, 28(06), 1, 24.
- Hoff, D.J. (2008d, December). Duncan confronts host of challenges at the Education Department. *Education Week*, 28(16), 22-25.
- Hoff, D.J. (2009, January). School struggling to meet key goal on accountability. *Education Week*, 28(16), 1, 14, 16.
- Kennedy-Manzo, K. (2008, June). Principals' group calls for national academic standards and tests. *Education Week*, 27(41), 6.

- Klein, A. (2008, June). Kennedy's illness raises doubts for NCLB. *Education, Week*, 27(41), 17-18.
- Kohn, A. (2001a, January). Fighting the tests: A practical guide to rescuing our schools. *Phi Delta Kappan*, Retrieved April 1, 2005, from <http://www.alfiekohn.org/articles.html>
- Kohn, A. (2001b, January). Beware of the standards, not just the tests. *Education Week*, Retrieved from <http://www.alfiekohn.org/teaching/edweek>
- 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.
- Lewis, A. (2005, July). *Research guidance: Assessment, accountability, action!* (CSE Report 658). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Linn, R. (2005, Summer). *Fixing the NCLB accountability system* (CRESST Policy Brief 8). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- 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.

- Marsh, J., & Pane, J., Hamilton, L. (2006). *Making sense of data-driven decision making in education*. (Occasional Paper), Rand Education. Retrieved February 04, 2009 from http://www.rand.org/pubs/occasional_papers/OP170/
- Marzano, R. (2006). *Classroom assessment and grading that works*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Maturation. (2009). In *Merriam-Webster Online Dictionary*. Retrieved February 10, 2009, from <http://www.merriam-webster.com/dictionary/maturation>
- McNeil, M. (2008, September). States cite capacity gap in aid for schools on NCLB. *Education Week*, 28(5), 10.
- Medical University of South Carolina. (n.d.) *Statistical Significance*. Retrieved February 19, 2009 from <http://www.musc.edu/dc/icrebm/statistical/significance.html>
- Miles K. H. (2001, September). Putting money where it matters. *Educational Leadership*, 59(1), 53-57.
- National Office for Research Measurement and Evaluation Systems (NORMES). (2009.) *School performance report*. Retrieved February 19, 2009 from <http://normessasweb.uark.edu/src08/>
- 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.
- Olson, A. (2007, January). Growth measures for systemic change. *The School Administrator*, 64(1), 10-16.

- O.U.R. Cooperative (n.d.). *Total instructional alignment*. Retrieved February 24, 2009, from http://www.oursc.k12.ar.us/total_instructional_alignment.html
- Pascopella, A. (2006, January). Nitty-gritty data. *The District Administrator*, 42(1), 36-41.
- Pascopella, A. (2008, February). State of the Superintendency. *The District Administrator*, 44(2), 32-36.
- Perie, M., Marion, S., & Gong, B. (2007, February). *A framework for considering interim assessments*. Dover, NH: The National Center for the Improvement of Educational Assessment, Inc., Retrieved October 8, 2008, from <http://www.nciea.org>
- Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system* (Policy Brief). Aspen CO: The Aspen Institute, Center for Assessment.
- 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. (2005a, September). AYP wriggle room running out, *Education Leadership*, 63(1), 85-86.
- Popham, J.W. (2005b, November). Can growth ever be beside the point? *Education Leadership*, 63(3), 83-84.
- Popham, J.W. (2006, February). Assessment for leaning: An endangered species? *Education Leadership*, 63(5), 82-83.

- Popham, J.W. (2007a, March). Another bite out of the apple. *Education Leadership*, 64(6), 83-84.
- Popham, J.W. (2007b). *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. (2005, 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.
- Renaissance Learning (2006). *With the right information you can help your students shine like stars*. [Brochure]. Renaissance Learning Inc. Wisconsin Rapids, WI.
- Runyon, R.P, Coleman, K.A., & Pittenger, D.J. (2000). *Fundamentals of behavioral statistics*. Boston, MA: McGraw Hill.

- 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.
- Schmoker, M. (2006). *Results Now: How we can achieve unprecedented improvements in teaching and learning*. Alexandria, VA: Association for Supervision and Curriculum Development.
- 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*. (Occasional Paper), Retrieved September 26, 007 from <http://www.rand.org/publications/randreview>.
- 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.

Toch, T. (2008, October). Salvaging accountability. *Education Week*, 28(06), 30-31, 36.

Tomlinson, C. (2008, January). Leaning to love assessment. *Education Leadership*, 65(4), 8-13.

Triangulation in Education Research. (n.d.). Retrieved on March 9, 2009, from <http://www.geocities.com/zulkardi/submit3.html>

Trochim, W. M. (2008). *The research knowledge base*. (2nd Edition). Retrieved January 09, 2009 <http://www.socialresearchmethods.net/kb/>

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 (DOE). (2003, February). *Standards and assessments*. Title I Director's Conference, Retrieved September 5, 2006, from <http://www.ed.gov>

United States Department of Education Institute of Education Sciences. (2008, December). *WWC procedures and standards handbook*. Version 2.0: IES What Works Clearinghouse, Retrieved February 9, 2009, from <http://ies.ed.gov/ncee/wwc/references/idocviewer/Doc.aspx?docId=19&tocId=6>

Wallace, D. (2000, February). Results, results, results? *Educational Leadership*, 57(5), 66-68.

Weiss, M. (2008, July). The growth model pilot isn't what you think it is. *Education Week*, 27(42), pp. 28-29.

- Wilcox, J. (2007, February). NCLB on the eve of reauthorization: Calls for fundamental overhaul greet new congress. *Education Update*, 49(2).
- Wisdom, S. (2008, December). *Predictors of academic success for high school students: The correlation between middle school Missouri Assessment Program scores and freshman year grade point average*. Unpublished doctoral dissertation, Lindenwood University, St. Charles, Missouri.
- Wong, K., Nicotera, A. (2007). *Successful schools and educational accountability: Concepts and skills to meet leadership challenges*. Boston, MA: Pearson Education Inc.
- Zehr, M. (2008, July). States struggle to meet achievement standards for ELLs. *Education, Week*, 27(43), 12.

APPENDIX A

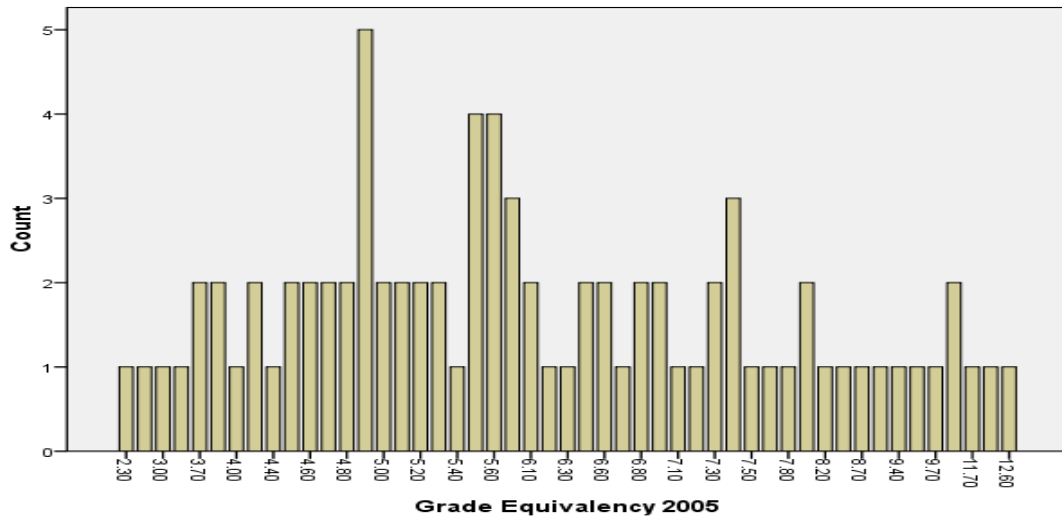


Figure A1. Frequency Distributions of Grade Equivalencies on the Sixth Grade STAR Math Pre-test 2005

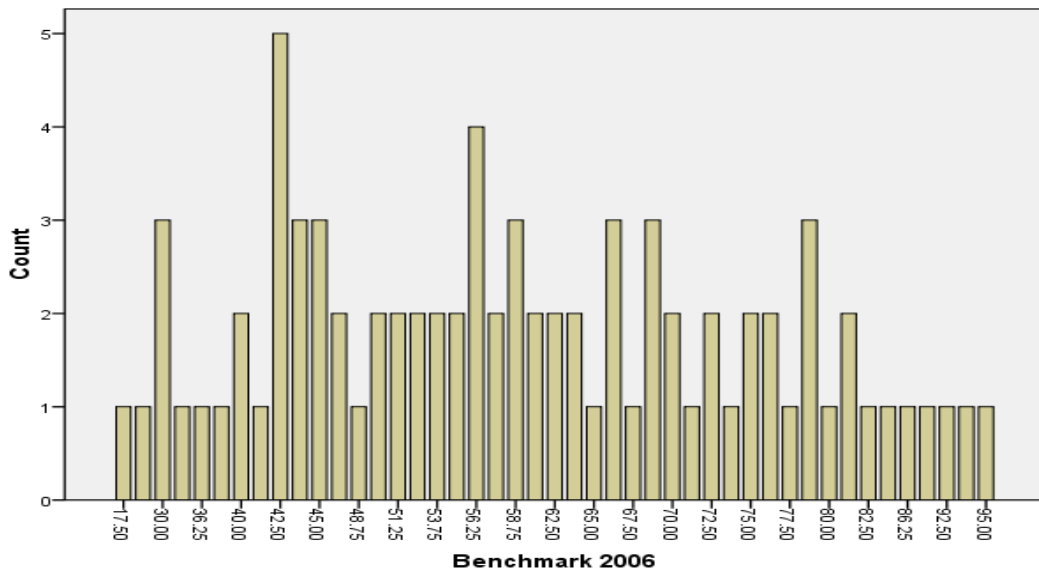


Figure A2. Frequency Distributions of Raw Score Percents on the Sixth Grade Benchmark Test 2006

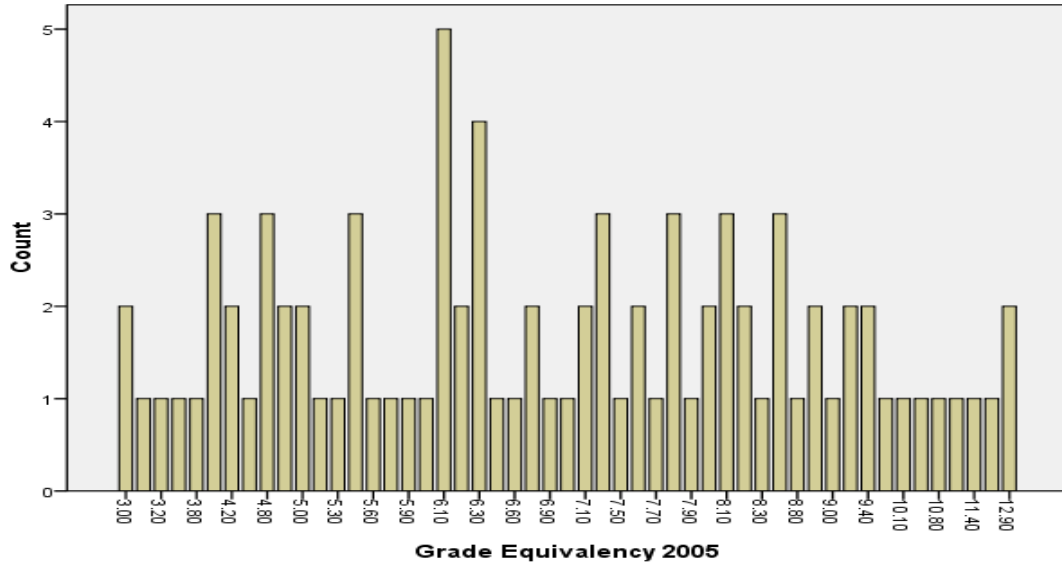


Figure A3. Frequency Distributions of Grade Equivalencies on the Seventh Grade STAR Math Pre-test 2005

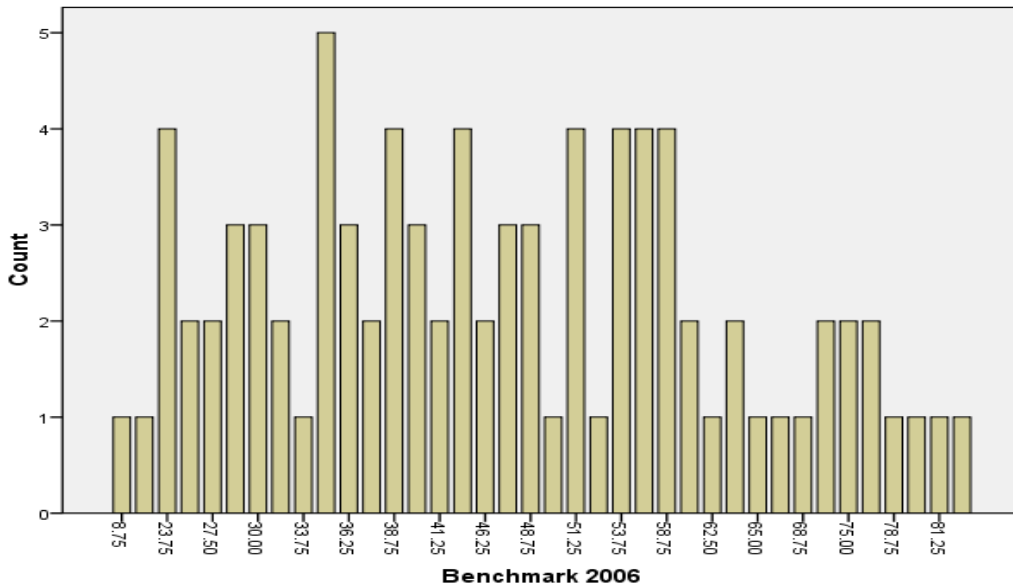


Figure A4. Frequency Distributions of Raw Score Percents on the Seventh Grade Benchmark Test 2006

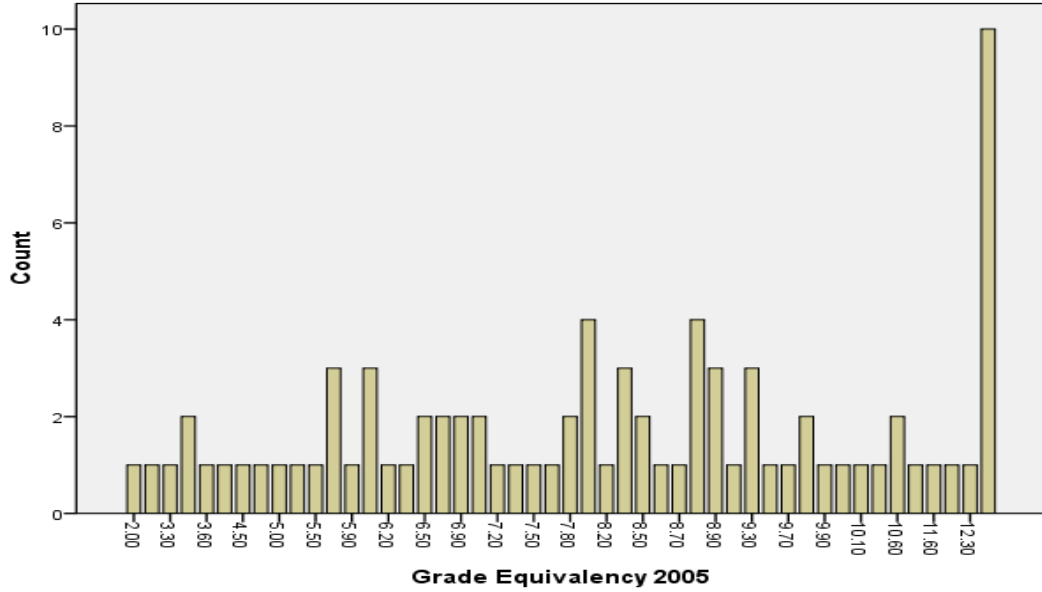


Figure A5. Frequency Distributions of Grade Equivalencies on the Eighth Grade STAR Math Pre-test 2005

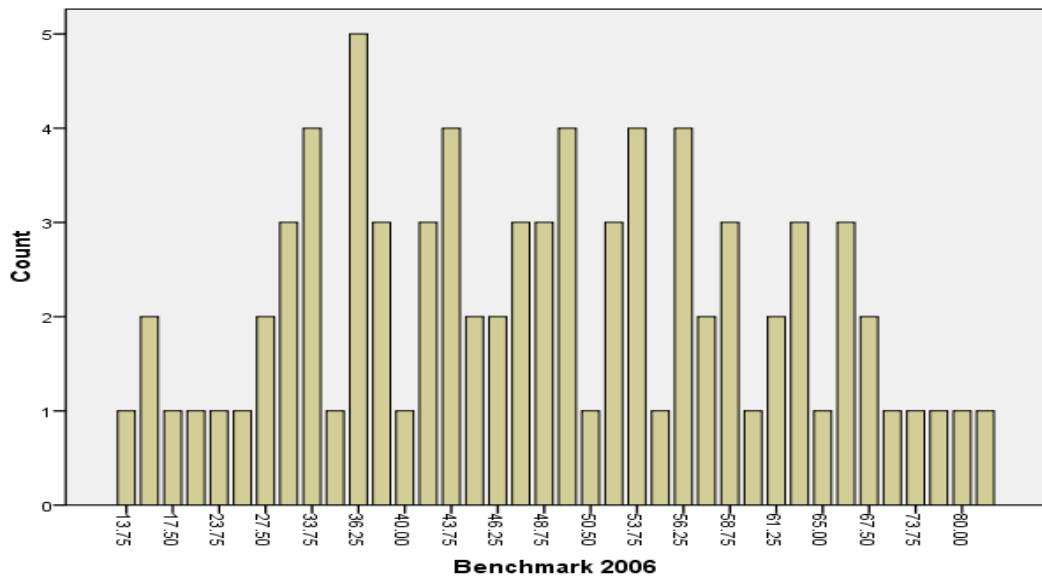


Figure A6. Frequency Distributions of Raw Score Percents on the Eighth Grade Benchmark Test 2006

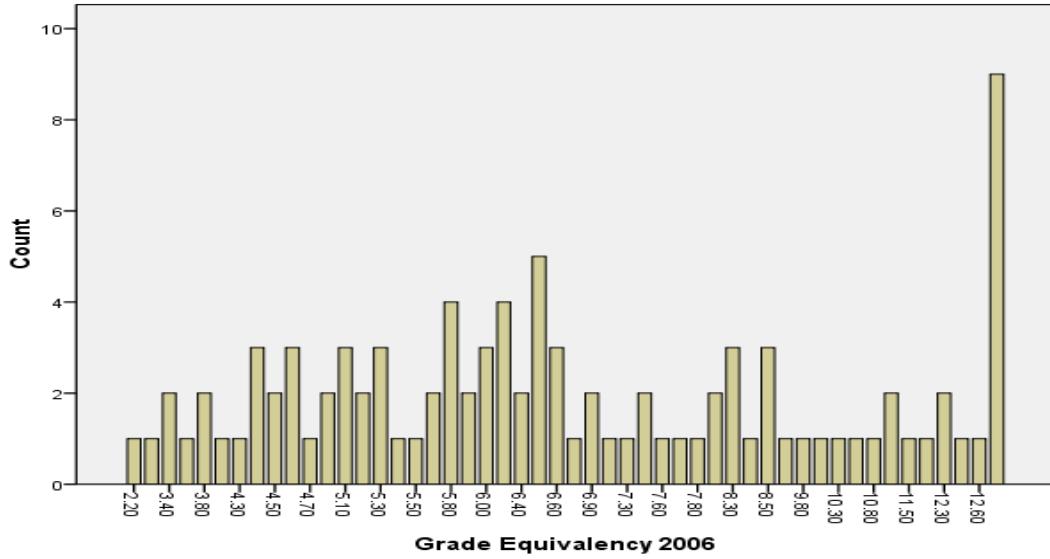


Figure A7. Frequency Distributions of Grade Equivalencies on the Sixth Grade STAR Math Pre-test 2006

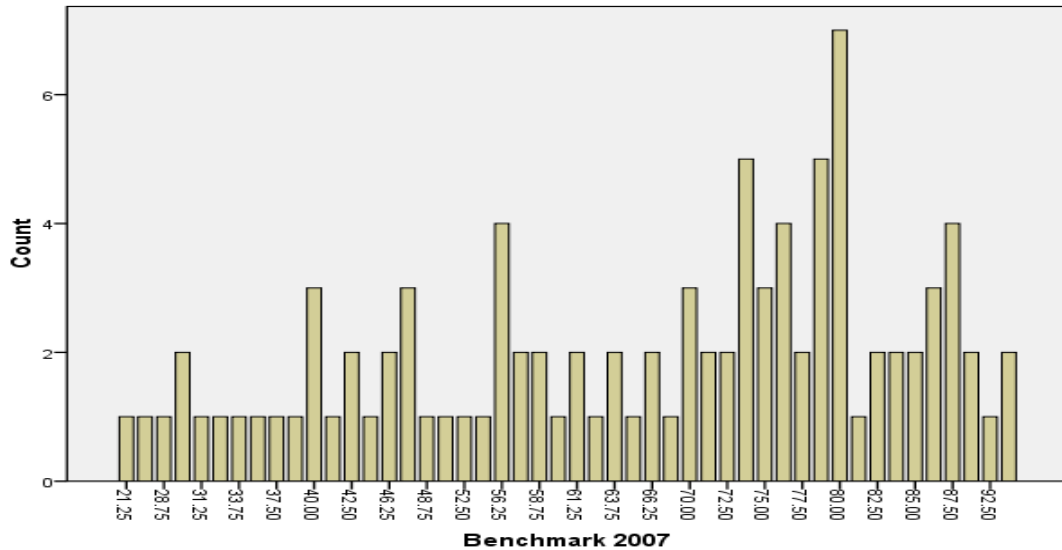


Figure A8. Frequency Distributions of Raw Score Percents on the Sixth Grade Benchmark Test 2007

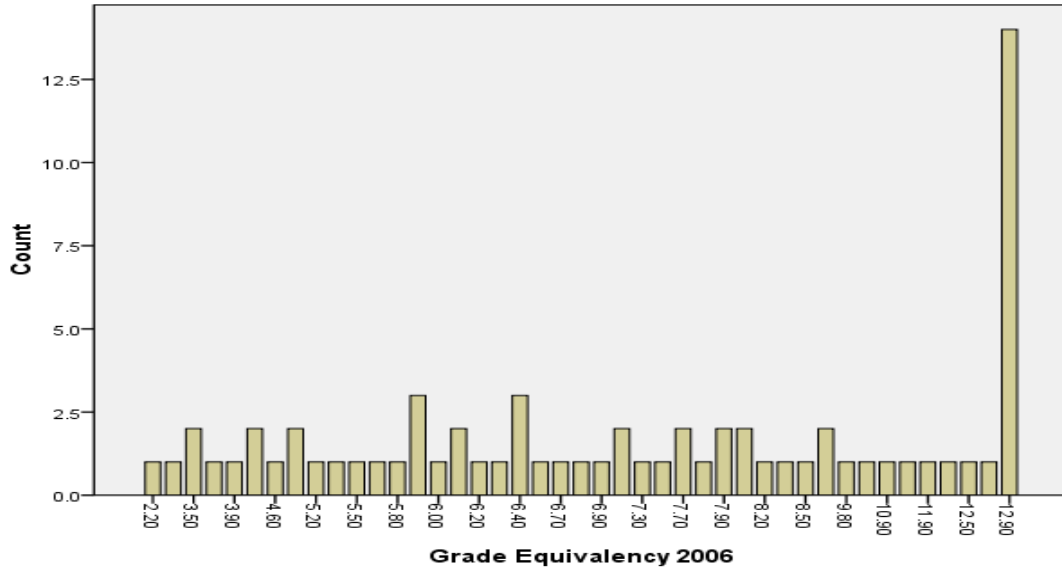


Figure A9. Frequency Distributions of Grade Equivalencies on the Seventh Grade STAR Math Pre-test 2006

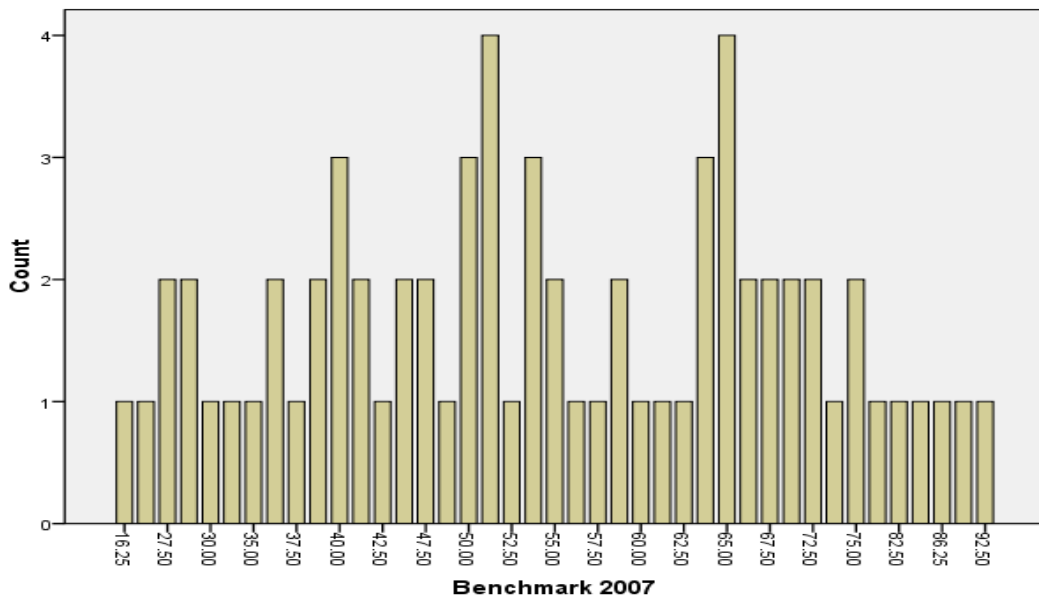


Figure A10. Frequency Distributions of Raw Score Percents on the Seventh Grade Benchmark Test 2007

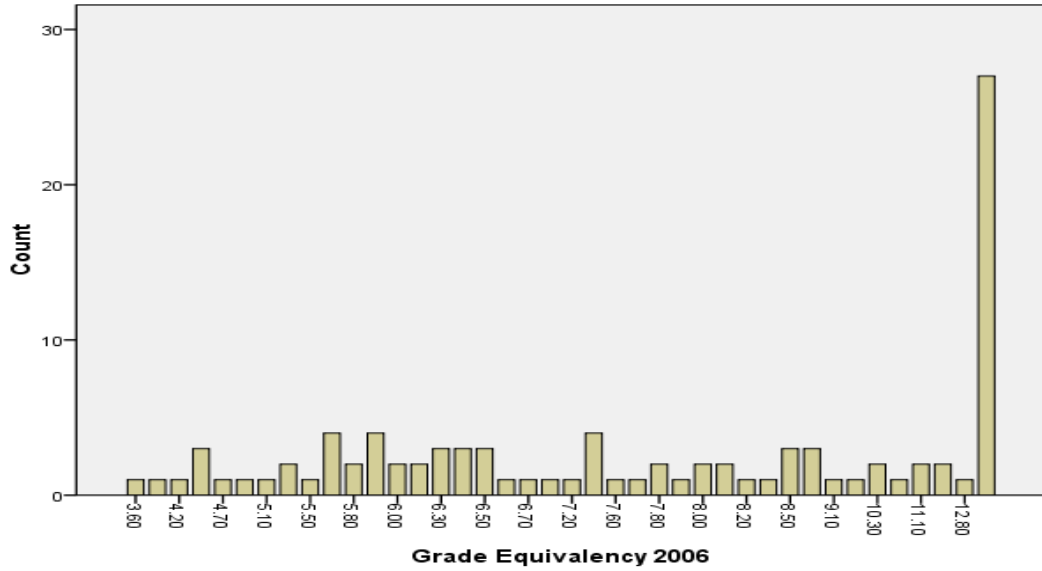


Figure A11. Frequency Distributions of Grade Equivalencies on the Eighth Grade STAR Math Pre-test 2006

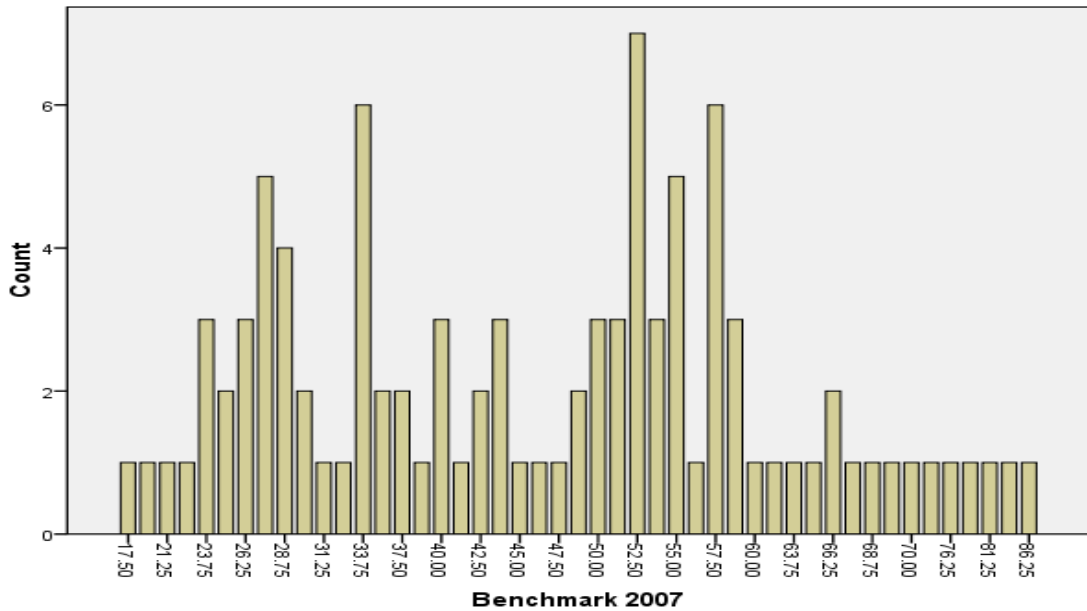


Figure A12. Frequency Distributions of Raw Score Percents on the Eighth Grade Benchmark Test 2007

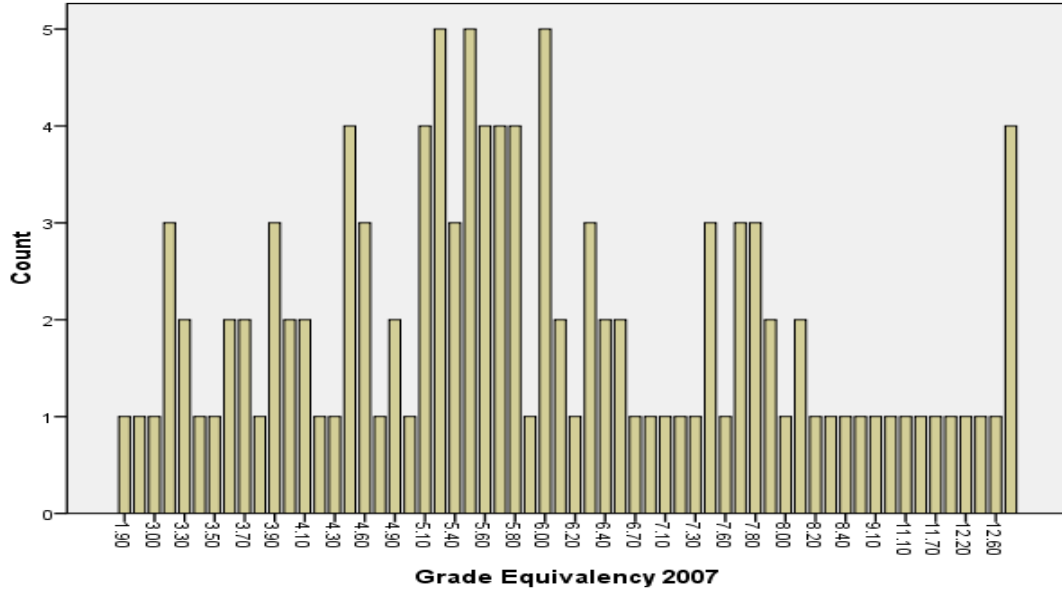


Figure A13. Frequency Distributions of Grade Equivalencies on the Sixth Grade STAR Math Pre-test 2007

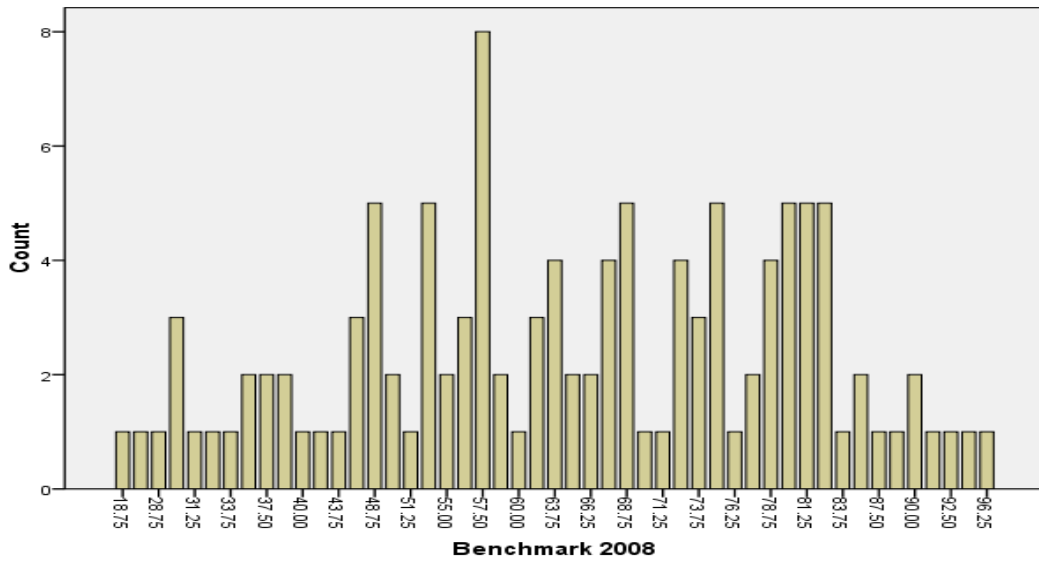


Figure A14. Frequency Distributions of Raw Score Percents on the Sixth Grade Benchmark Test 2008

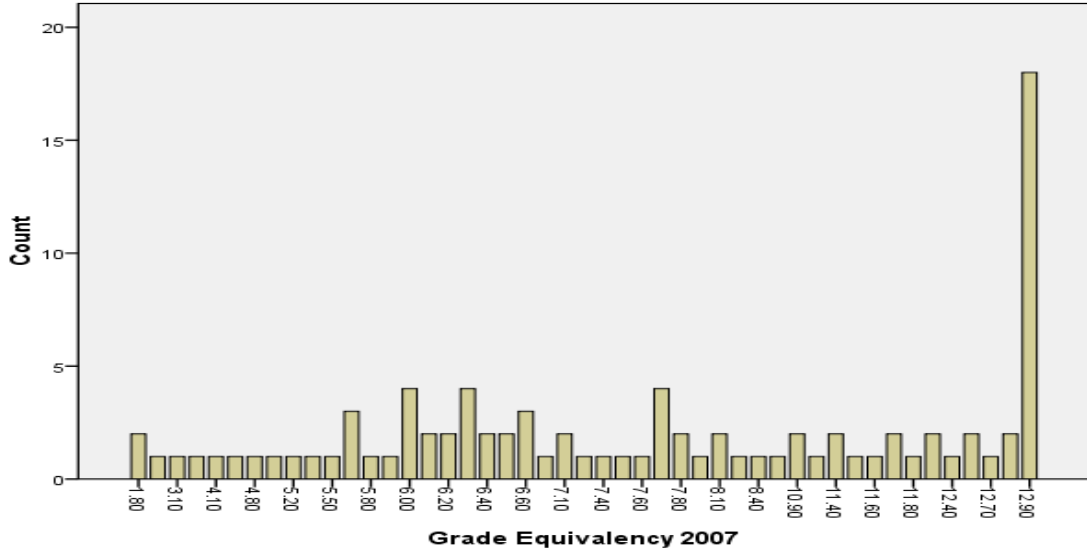


Figure A15. Frequency Distributions of Grade Equivalencies on the Seventh Grade STAR Math Pre-test 2007

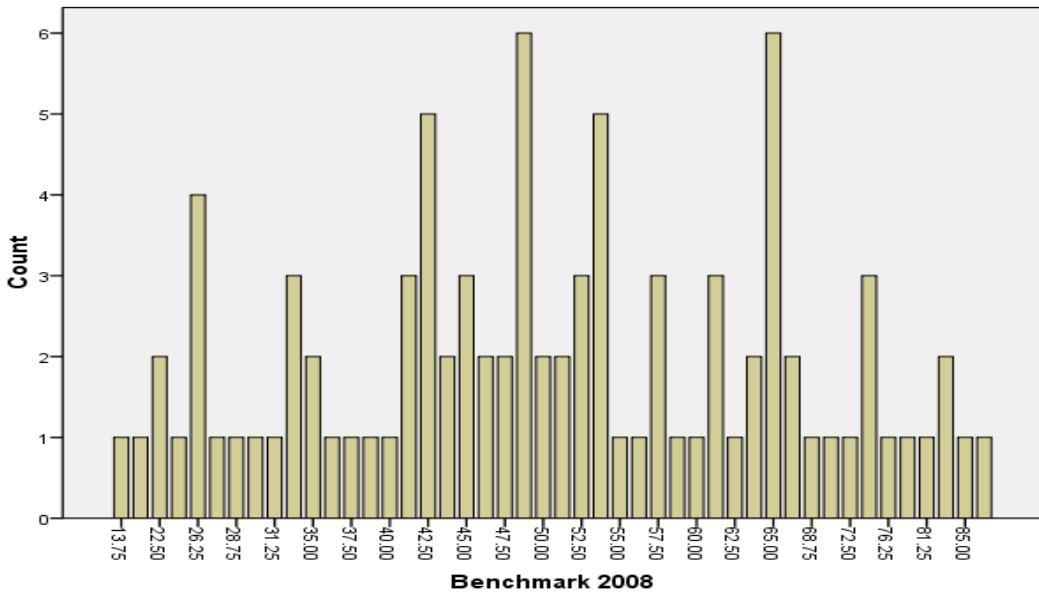


Figure A16. Frequency Distributions of Raw Score Percents on the Seventh Grade Benchmark Test 2008

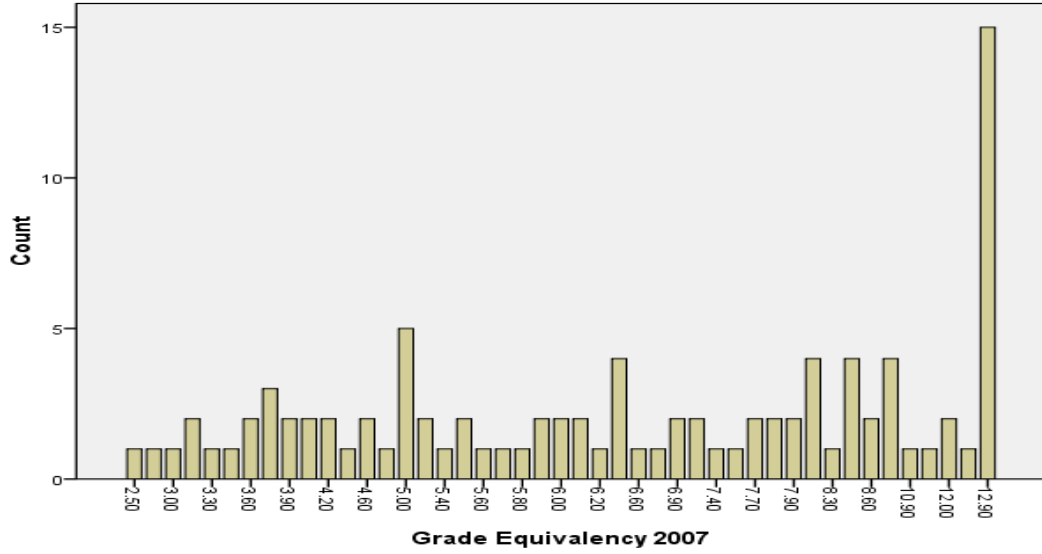


Figure A17. Frequency Distributions of Grade Equivalencies on the Eighth Grade STAR Math Pre-test 2007

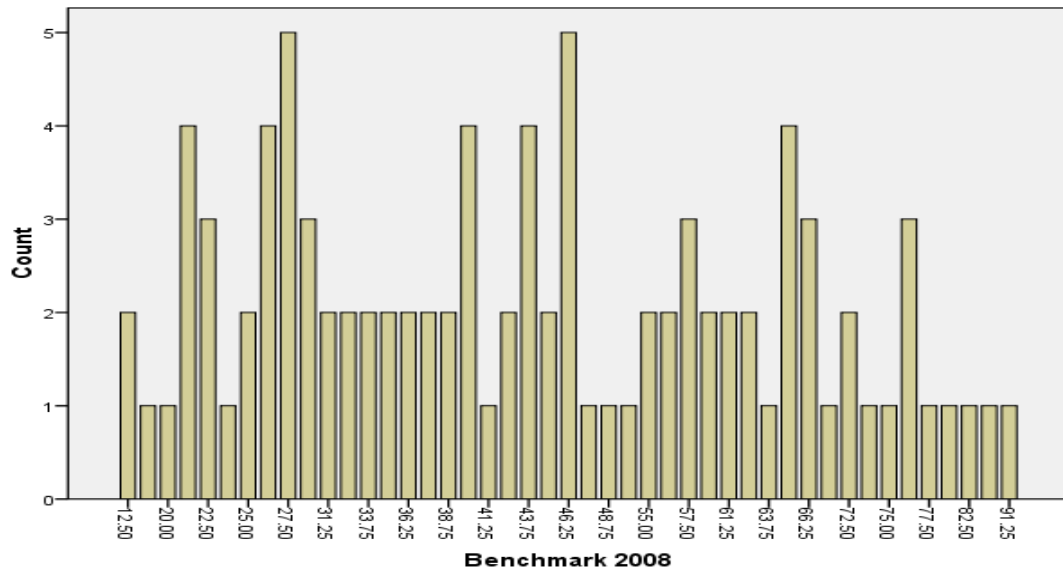


Figure A18. Frequency Distributions of Raw Score Percents on the Eighth Grade Benchmark Test 2008

APPENDIX B

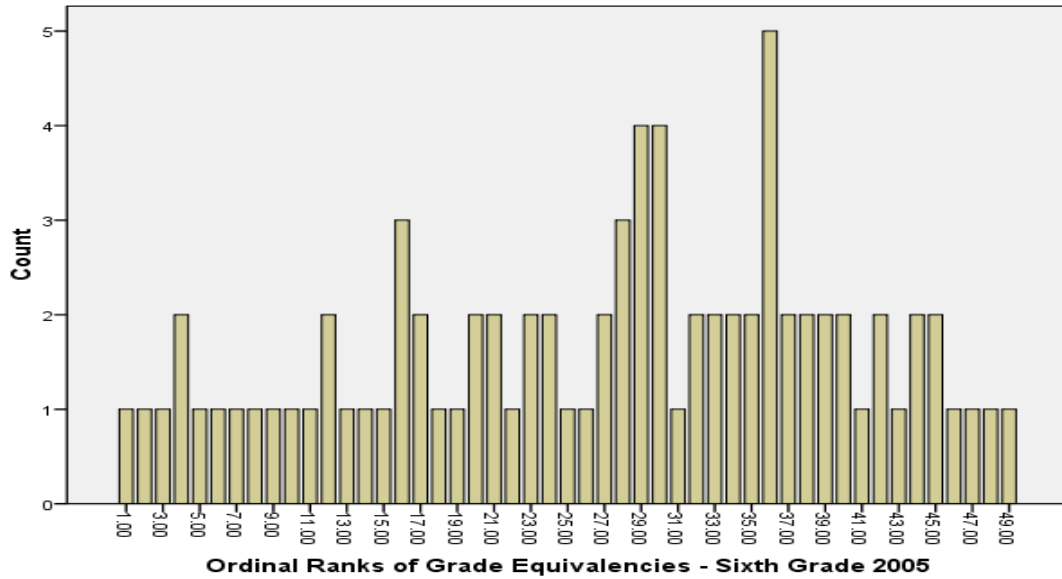


Figure B1. Frequency Distributions of the Ordinal Ranks for the Grade Equivalencies on the Sixth Grade STAR Math Pre-test 2005

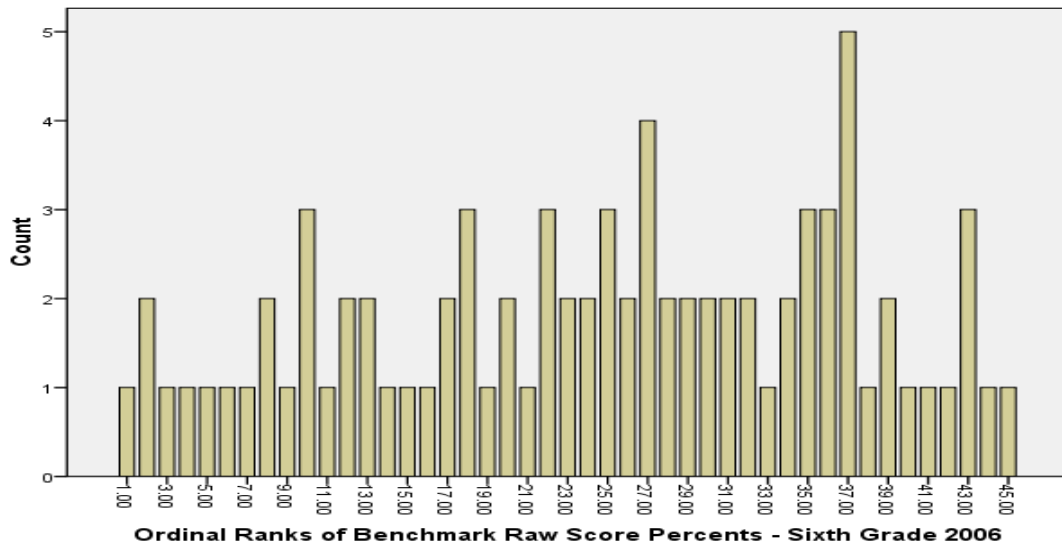


Figure B2. Frequency Distributions of the Ordinal Ranks for the Raw Score Percents on the Sixth Grade Benchmark Test 2006

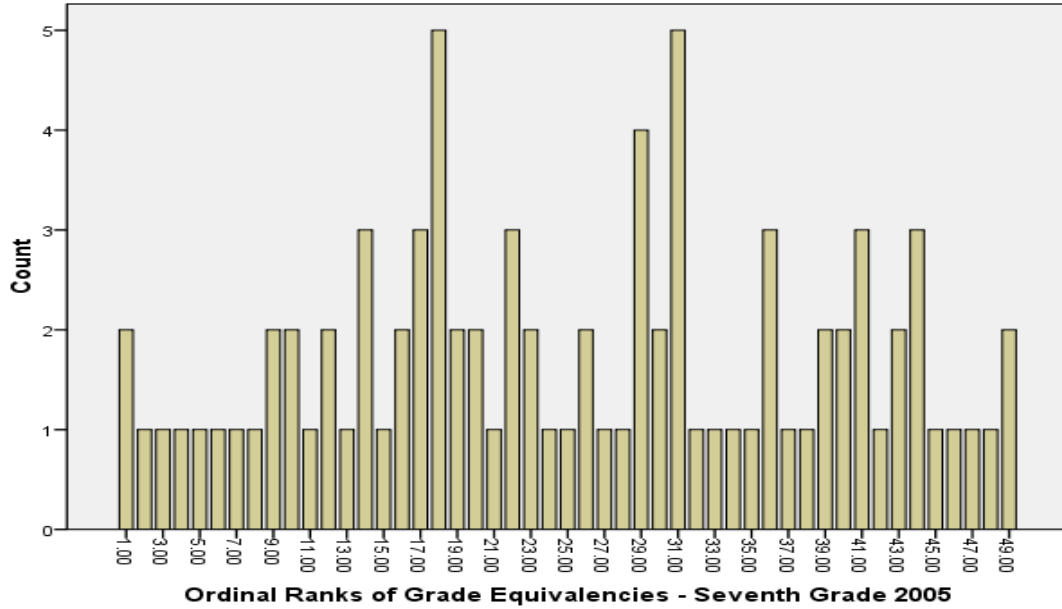


Figure B3. Frequency Distributions of the Ordinal Ranks for the Grade Equivalencies on the Seventh Grade STAR Math Pre-test 2005

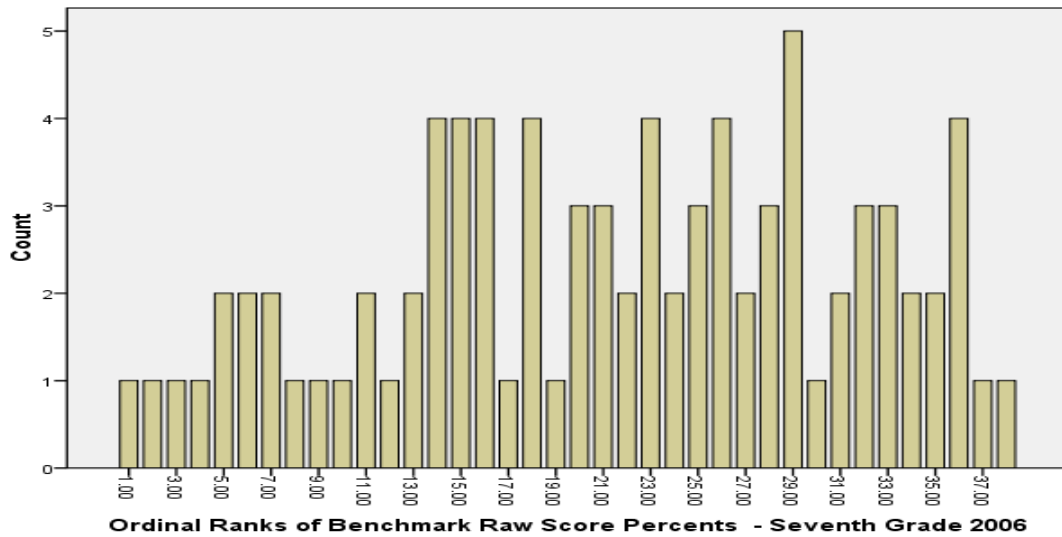


Figure B4. Frequency Distributions of the Ordinal Ranks for the Raw Score Percents on the Seventh Grade Benchmark Test 2006

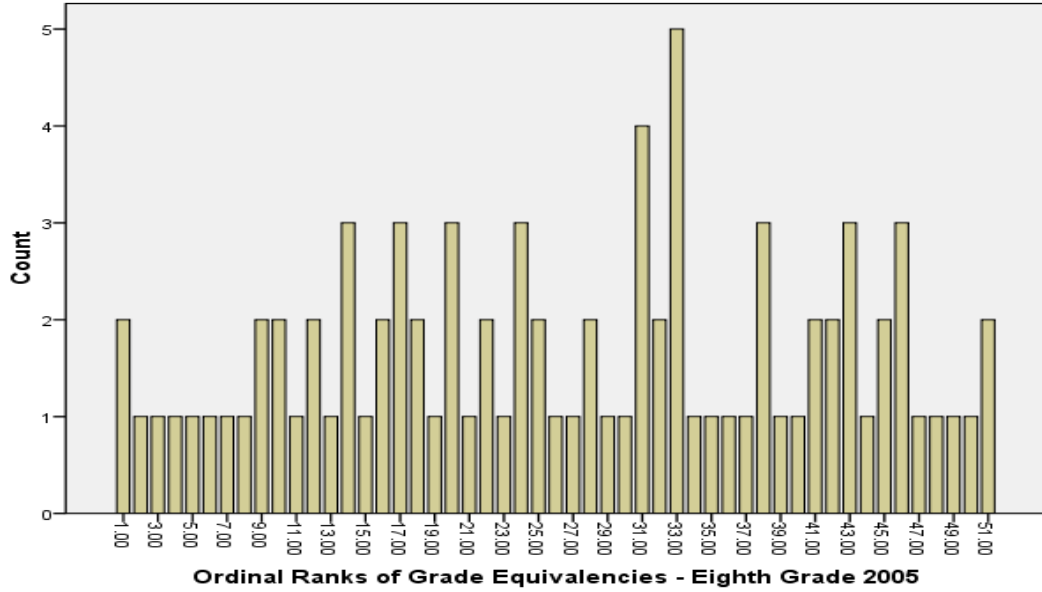


Figure B5. Frequency Distributions of the Ordinal Ranks for the Grade Equivalencies on the Eighth Grade STAR Math Pre-test 2005

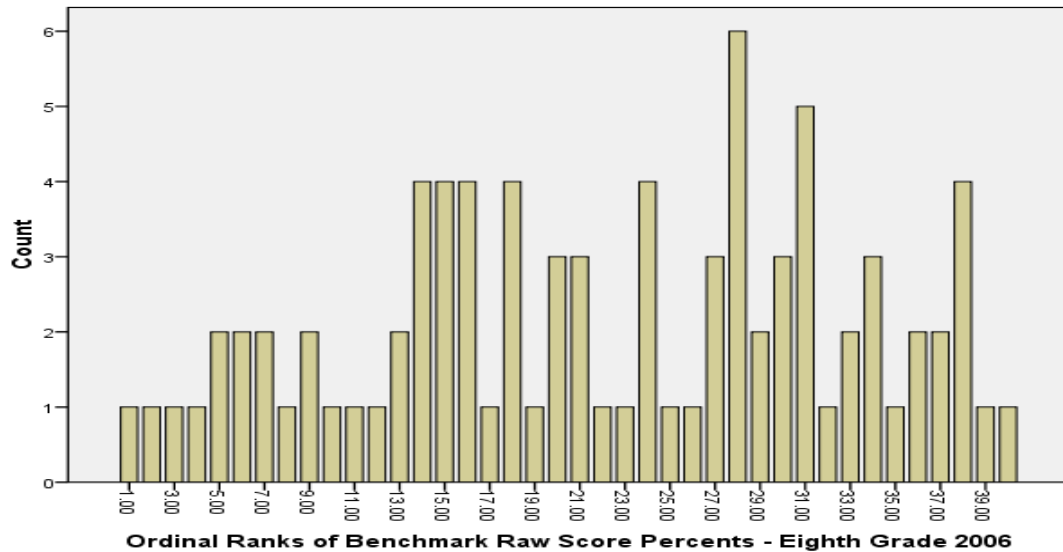


Figure B6. Frequency Distributions of the Ordinal Ranks for the Raw Score Percents on the Eighth Grade Benchmark Test 2006

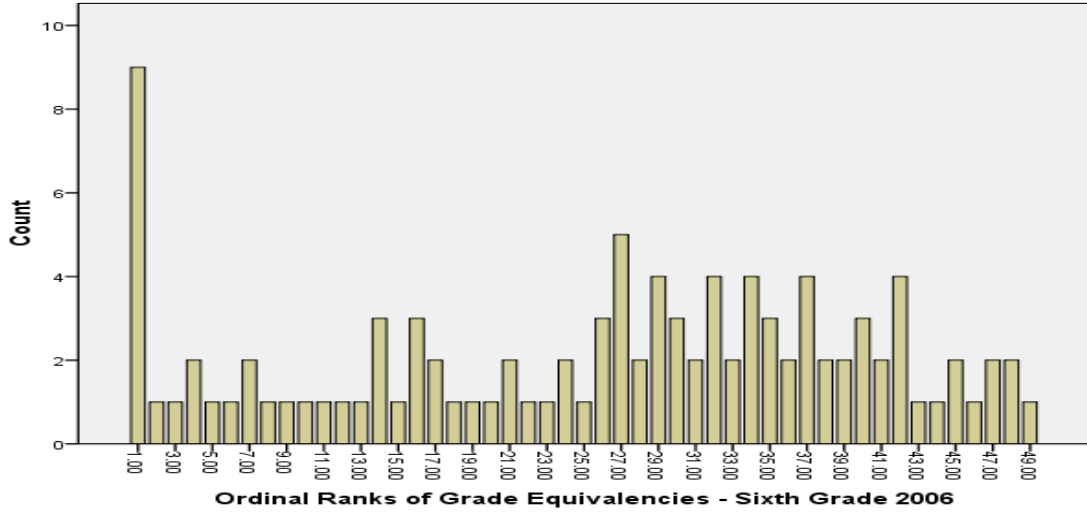


Figure B7. Frequency Distributions of the Ordinal Ranks for the Grade Equivalencies on the Sixth Grade STAR Math Pre-test 2006

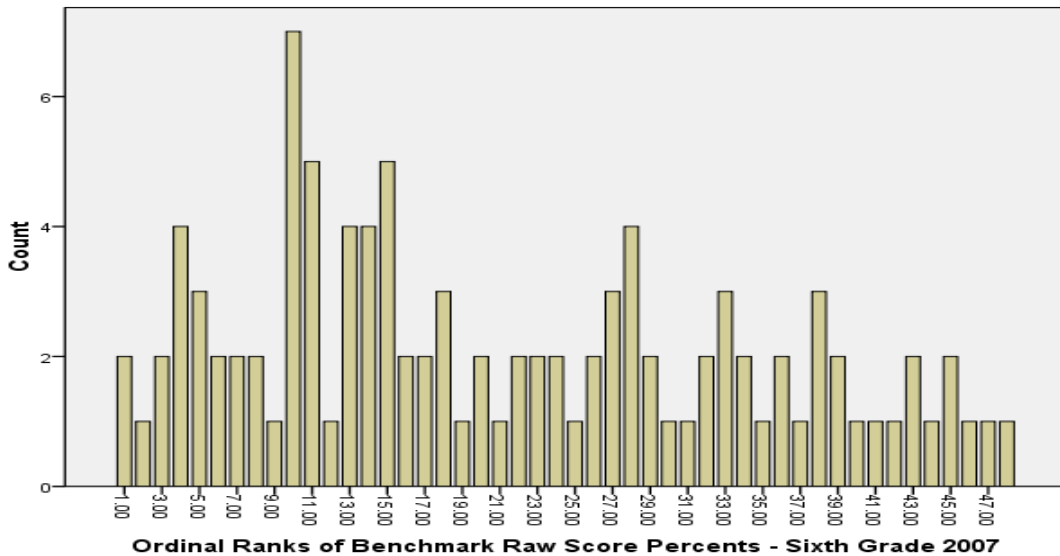


Figure B8. Frequency Distributions of the Ordinal Ranks for the Raw Score Percents on the Sixth Grade Benchmark Test 2007

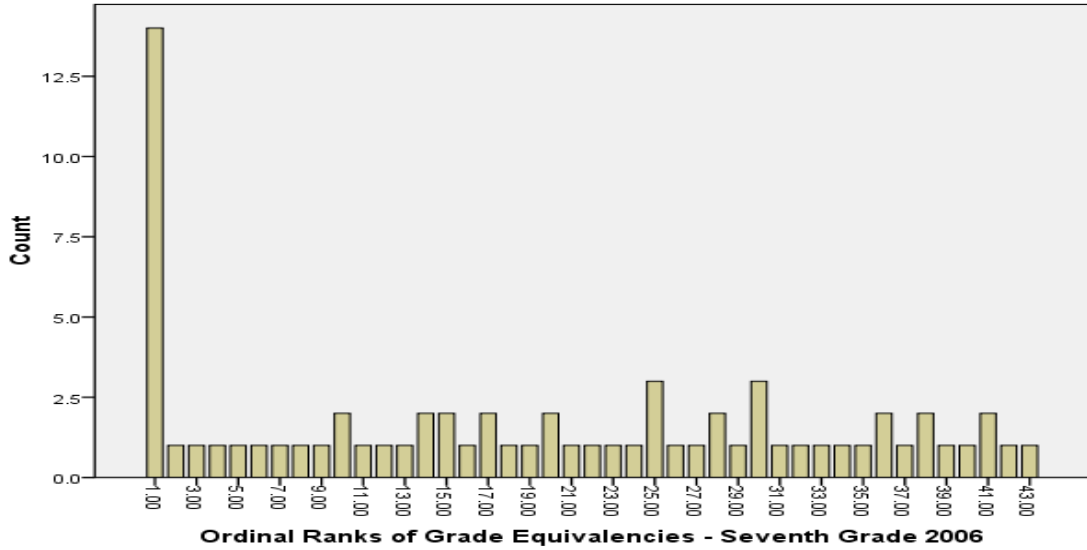


Figure B9. Frequency Distributions of the Ordinal Ranks for the Grade Equivalencies on the Seventh Grade STAR Math Pre-test 2006

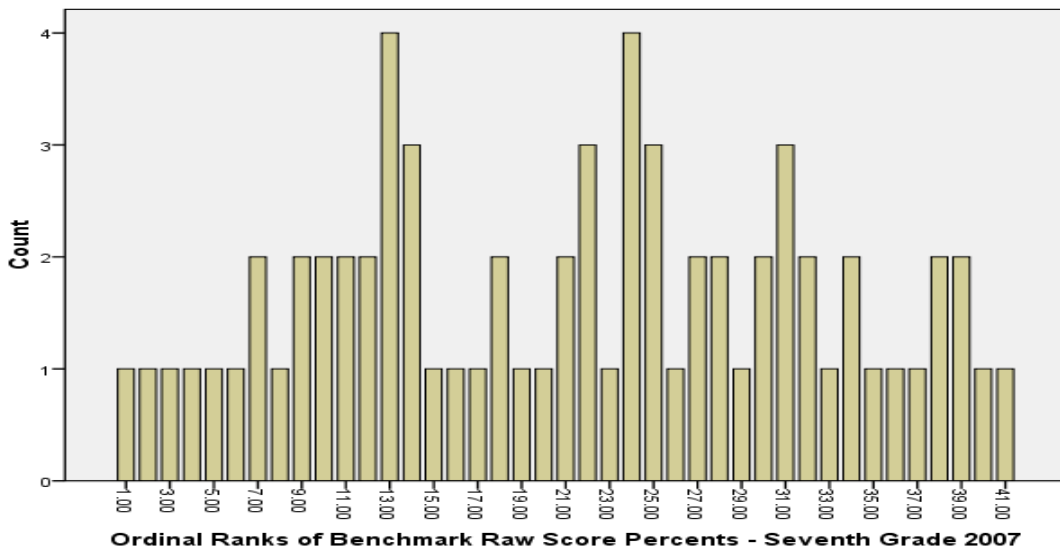


Figure B10. Frequency Distributions of the Ordinal Ranks for the Raw Score Percents on the Seventh Grade Benchmark Test 2007

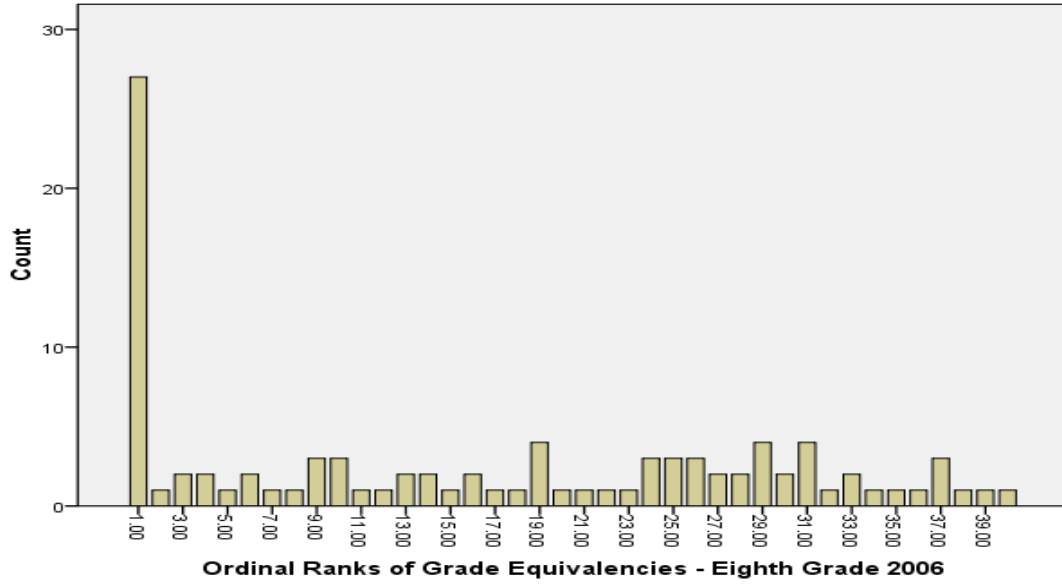


Figure B11. Frequency Distributions of the Ordinal Ranks for the Grade Equivalencies on the Eighth Grade STAR Math Pre-test 2006

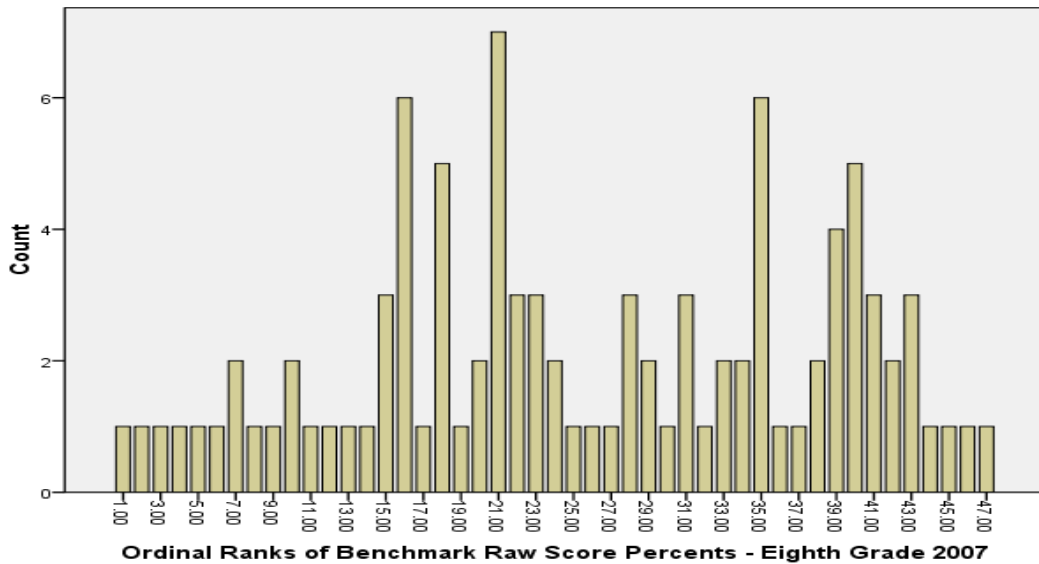


Figure B12. Frequency Distributions of the Ordinal Ranks for the Raw Score Percents on the Eighth Grade Benchmark Test 2007

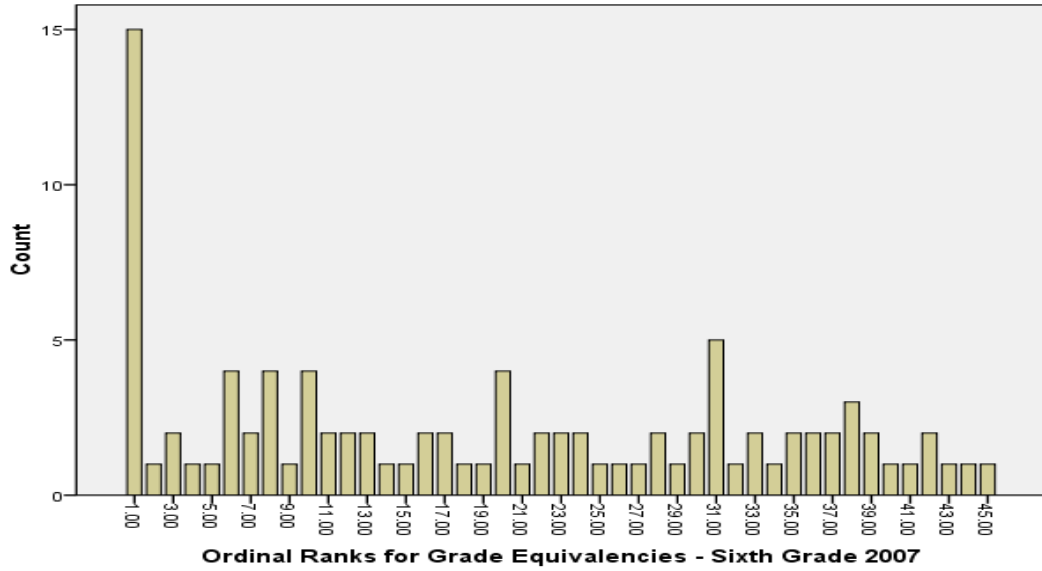


Figure B13. Frequency Distributions of the Ordinal Ranks for the Grade Equivalencies on the Sixth Grade STAR Math Pre-test 2007

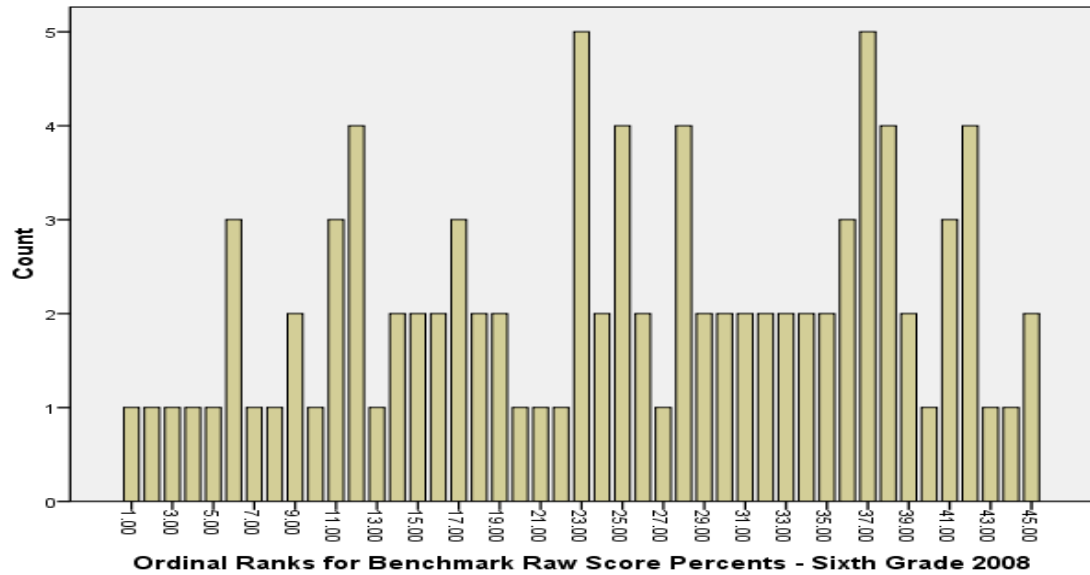


Figure B14. Frequency Distributions of the Ordinal Ranks for the Raw Score Percents on the Sixth Grade Benchmark Test 2008

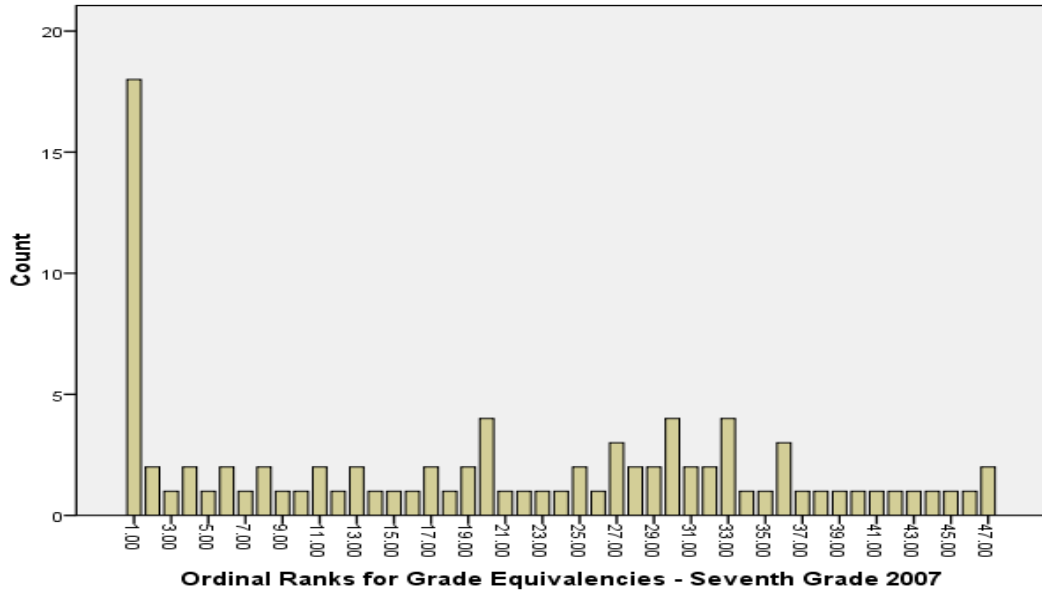


Figure B15. Frequency Distributions of the Ordinal Ranks for the Grade Equivalencies on the Seventh Grade STAR Math Pre-test 2007

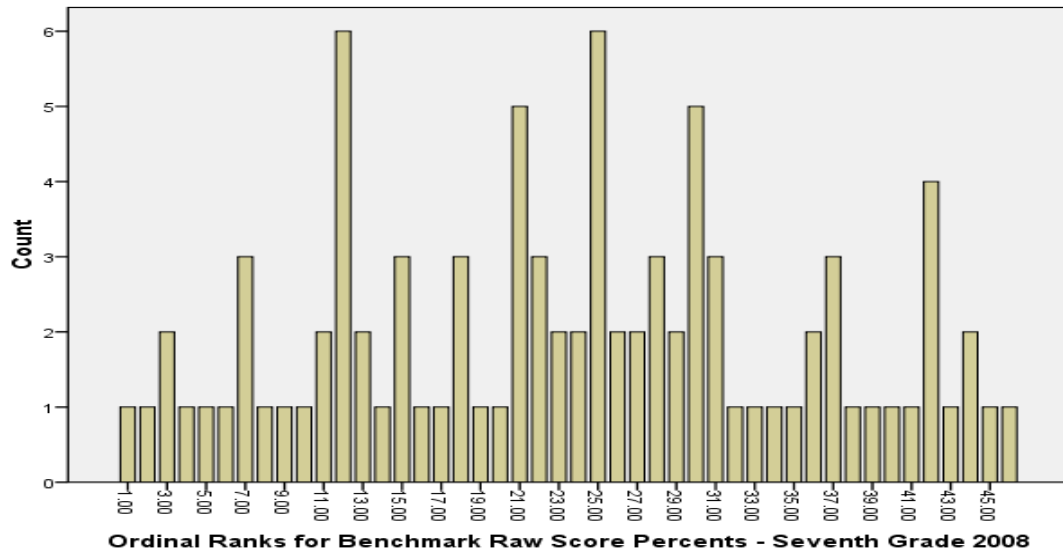


Figure B16. Frequency Distributions of the Ordinal Ranks for the Raw Score Percents on the Seventh Grade Benchmark Test 2008

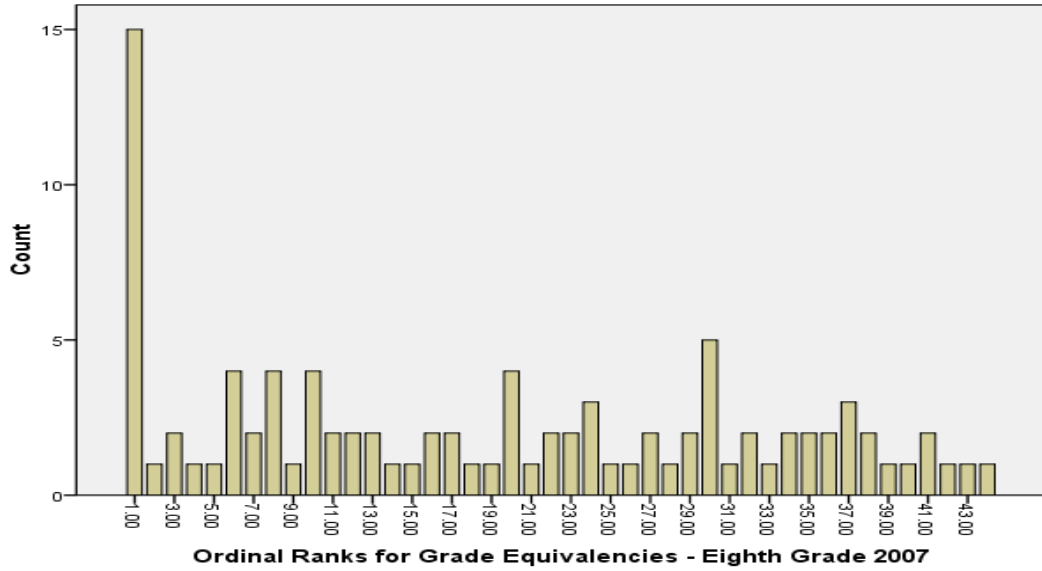


Figure B17. Frequency Distributions of the Ordinal Ranks for the Grade Equivalencies on the Eighth Grade STAR Math Pre-test 2007

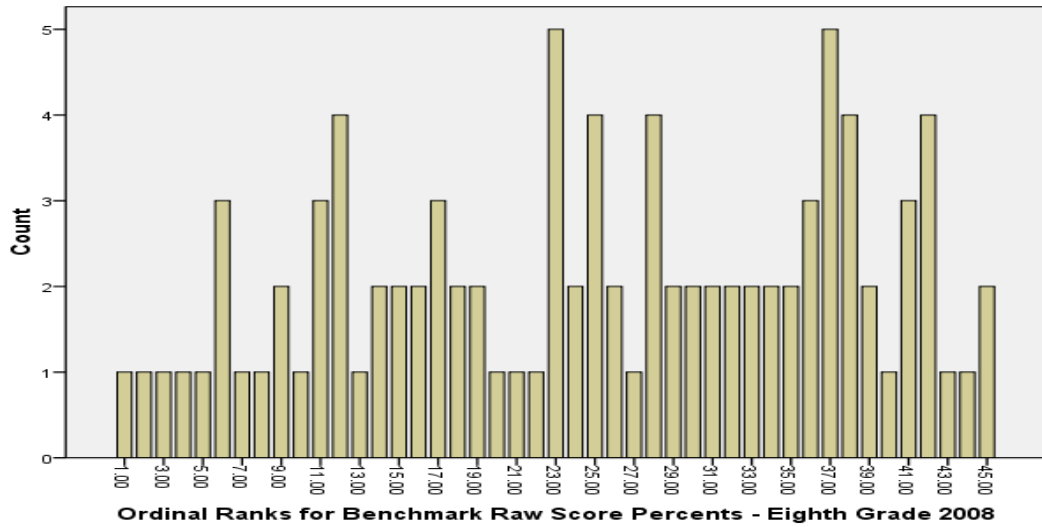


Figure B18. Frequency Distributions of the Ordinal Ranks for the Raw Score Percents on the Eighth Grade Benchmark Test 2008

APPENDIX C

E-mail to Arkansas Administrators

I am asking you to respond to the following survey concerning pre-assessments and your views on their ability to predict achievement on the Arkansas Benchmark Test. 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 reply to the attached survey and return it to me at pconner@bobcat.k12.ar.us. Your help and effort is appreciated. Please call me at 1-870-423-3313 or email me at the above address if you have any questions.

Sincerely,

Patricia Conner
District Testing Coordinator
Berryville Schools
902 W. Trimble
Berryville, AR 72616

Survey of Arkansas Educators

Survey

Please respond by placing an X in front of the number that most appropriately answers the questions below.

Please check which building you work in:

_____ Elementary _____ Intermediate
_____ Middle School _____ High School

1. Does your school use grade-level pre-assessments at the beginning of the year in math to determine student achievement levels?

_____ YES _____ NO

2. Does your school use the STAR Math Test from the Renaissance Learning Company?

_____ YES _____ NO

3. Do you believe pre-assessments accurately provide a predictor for student achievement on the Arkansas Benchmark Test? (Please answer whether not your district uses a pre-assessment test.)

_____ YES _____ NO

VITA

Patricia A Conner is currently the District Test Coordinator for Berryville Public Schools, in Berryville, Arkansas. Teaching experiences have included grades 7-12 social studies and college level education courses. She has also served the district as the Director of the Alternative Education Program. Specific areas of interest are curriculum and instruction and assessment with data management.

Educational studies have resulted in an Education Specialist Degree in educational leadership from Lindenwood University, a Master of Education in educational leadership from the University of Arkansas, and a Bachelor of Science in Education Degree from College of the Ozarks.