State of Nebraska Reading Comprehension/Vocabulary Test and the Local American Literature Criterion-referenced Test for Public School Eleventh Grade Students: A Correlational Study

A dissertation submitted

by

Priscilla A. Beckmann

to

College of St. Mary

In partial fulfillment of the requirement

For the degree of

DOCTOR IN EDUCATION

With an emphasis on

Education Leadership

This dissertation has been accepted for the faculty of

College of Saint Mary by:

Dr. Lois Linden, Ed.D., R.N.

Dr. Patricia Morin, PhD., R.N.

Dr. Daniel Cox, Ph.D., Curriculum and Instruction

We hereby certify that this dissertation, submitted by Priscilla A. Beckmann, conforms to acceptable standards and fully fulfills the dissertation requirements for the degree of Doctor in Education from College of Saint Mary

Dr. Lois Linden, Ed.D., R.N.
Dissertation Committee Chair

Dr. Patricia Morin, Ph. D., R.N,
Dissertation Committee Member

Dr. Daniel Cox, Ph.D., Curriculum and Instruction
Dissertation Committee Member

"We ought to act in such a way that what is true can be verified to be so."

Jacob Bronowski, 1965

This dissertation is dedicated to my father, who did not finish high school, and my mother, who finished her B.S. at age 53, after 20 years of part-time education and 36 years of full-time work, for their sincere belief that education is important in developing the self and preserving the society. In addition, I would like to thank my husband for his support of me in this endeavor–my life-long dream. And, just as importantly I would like to thank my husband and our grown children and their families for focusing on those same principles of education, personal growth, and social justice.

In addition, this paper is offered as the manifestation of respect I have for all the educators who I have learned from, who I have worked with, and who have shown and continue to show professionalism, integrity, and commitment in their teaching. These educators are the inspiration for my vocation, and they are the voice of my conscience reminding me every day to demonstrate this same professionalism, integrity, and commitment in my teaching.

ACKNOWLEDGEMENTS

Thank you to the many people who have made this research possible and useful.

Specifically, I would like to thank my Doctoral Committee: Dr. Lois Linden, Ed.D., R.N.; Dr. Patricia Morin, Ph.D., R.N.; Dr. Daniel Cox, Ph.D, Curriculum and Instruction. All of your insight, patience, and encouragement facilitated the completion of this project.

Thank you Dr. Steve Sexton, Ed. D., Education Administration, Superintendent of Schools, for your permission to do the research and your trust that the results will be useful to the district.

Thank you Mr. Terry Snyder, Executive Curriculum Director; for suggesting this particular research and enthusiastically supporting me as the year went along.

Thank you Mr. Joe Sajevic, M. Ed., Education Administration, Senior High School Principal, for endorsing my journey to this degree.

Thank you to Ms. Deb Horrocks, Curriculum Office Assistant; Ms. Kelly Monke, Curriculum Office Assistant; and Ms. Freda Hocking, Curriculum Office Assistant, Emeritus, for all your work in providing the records for this research.

Thank you to my friends who are also doctoral students with me, some who are finished and some who still are in the process, specifically, Shannon Hansen, Sheli Hensely, Dr. Brenda Schiermeyer, Stacey Smith, and Dr. Mary Trehearn. You made this trek less a process and more an adventure.

And finally, thank you, my colleagues, Mr. Justin Bigsby, Senior High English Department Chair and American Literature Teacher; Ms. Stacey Smith, American Literature Teacher; Ms. Shannon Hansen, American Literature Teacher; Dr. Daniel Cox, American Literature Teacher, and Mr. Steve Franks, American Literature Teacher for your assistance without argument.

Table of Contents

Abstract

This study focused on the effectiveness of standards-based testing and was designed to determine the relationship of student scores on the locally constructed, standards-based, criterion-referenced test (CRT) for approximately 280 eleventh grade students at a Midwestern, Class III district, operating a Class A high school, and the scores of those same students on the Nebraska State Reading Comprehension/vocabulary Proficiency Examination for eleventh grade, public high school students. This study determined whether the district created tests showed a positive correlation between the quarter exams in their entirety with the Nebraska Reading Comprehension/vocabulary Proficiency Examination, and determined whether a positive correlation existed between the cold-read portion of the district created quarter CRT and the Nebraska Reading Comprehension/vocabulary Proficiency Examination. Correlations showed the positive and to be of .50 or higher between student scores of the eleventh grade students who had taken both Quarter A and Quarter B tests and the Nebraska Reading Comprehension/vocabulary Proficiency Examination in all sub-sets of students: first semester students Quarter A and Quarter B and second semester students Quarter A and Quarter B. In addition, the means for the male and female students on the Nebraska Reading Comprehension/vocabulary Proficiency Examination and the means for the Free and/or Reduced Lunch Program participants and non-participant students were examined and found to demonstrate a statistically significant correlation for each sub-set.

Chapter 1: Introduction

Background and Rationale

Performance expectations of public schools, demands on teacher competency, and demands on student learning have had a dramatic change in the United States since 2001. Since the elements and focus of the 2001 Elementary Secondary Education Act (2001 ESEA) were implemented into the public schools, the demand for reliable performance expectations has become more widespread, more comprehensive, and has had more accountability factors than reforms or changes in public school education at any previous era except at its inception around the time of the Civil War. The Homestead Act of 1862 encouraged development and earlier the federal government required that section 16 out of each 36 square mile township be set aside for public school.

> The national system of formal education in the United States developed in the 19th century. . . . until the 1840's the education system was highly localized and available only to wealthy people. Reformers who wanted all children to gain the benefits of education opposed this. . . . as a result of their efforts, free public education at the elementary level was available for all American children by the end of the 19th century (Thattai, D., 2001).

As a result of the 2001 ESEA, the "measuring" of student learning or student performance has become scrutinized more completely, and demands for scientific methodology to give credibility to these measurements are being addressed. In educational settings, however, results that show statistical reliability, and testing that contains content validity, as well as sufficiency of

concept measure have become a focus for the classroom teacher and curriculum director. No longer are school districts able to pronounce that its students who have graduated from high school are prepared; instead, school districts are expected to show proof of such student preparation through assessment of student knowledge. In Nebraska, until the 2009-2010 school year, the responsibility for this performance assessment and pronouncement fell to the local school district. Each Nebraska school district devised its own method of assessment based on Nebraska state guidelines and performance standards. Because of these national and state changes in performance expectations, a local, Midwestern, Class A school district, felt strongly that its students' learning needed to be in line with the state expectations. The district's commitment to learning was evident in its mission statement, and the sincerity of this belief is evidenced by its desire for measurable and meaningful accountability.

Research Question

In this Midwestern, Class III district, operating a Class A school, what was the relationship between the scores of the eleventh grade students on the Language Arts assessment, specifically American Literature, criterion-referenced (CRT) quarter tests devised by the public school district for these students, and the same students' scores on the test devised by the Nebraska for Reading Comprehension/vocabulary Proficiency Examination for eleventh grade students in the Nebraska?

And, in this Midwestern, Class III district, operating a Class A high school, what was the relationship between the locally devised, quarter reading

proficiency test (cold-read), and the same students' scores on the Nebraska Reading Comprehension/vocabulary Proficiency Examination devised by the Nebraska Education Department for all eleventh grade students in the Nebraska?

And, does a statistically significant difference occur between mean scores of males and females on the Nebraska Reading Comprehension/vocabulary Proficiency Examination?

And, does a statistically significant difference occur between mean scores of Free and/or Reduced Lunch Program participants and non-participants?

Sub-questions

Because this public school's student day was organized into the 4 X 4 block schedule, and, because of the gender distribution and the Free and/or Reduced Lunch Program student participants, questions arose regarding test preparation and outcomes for and on the Nebraska Reading Comprehension/vocabulary Proficiency Examination. Sub-questions arose for these topics:

1. Was there a difference in assessment scores for the semester Quarter A CRT, Quarter B CRT, and the Nebraska Reading Comprehensive/vocabulary Proficiency Examination measurement between those students who made up a group of approximately 50% of the eleventh grade students, who had completed both quarters of American Literature first semester before the state test was administered in the spring approximately 10 weeks after their formal Language Arts

instruction was completed, and those students who entered the American Literature class for the second semester but before the state test was administered in the spring approximately 10 weeks or fewer after their formal instruction began?

2. Was there a difference in assessment scores for Quarter A cold-read, Quarter B cold-read, and the Nebraska Reading Comprehensive/vocabulary Proficiency Examination measurement between those students who made up a group of approximately 50% of the eleventh grade students, who had completed both quarters of American Literature first semester before the state test was administered in the spring approximately 10 weeks after their formal language arts instruction was completed, and those students who entered the American Literature class for the second semester but before the state test was administered in the spring approximately 10 weeks or fewer after their formal instruction began?

3. Was there a difference in the mean scores for the groups related to the particular area of measurement between those students who made up the male student grouping of the eleventh grade, approximately, 50% of the eleventh grade students, and those students who made up the female student grouping of the eleventh grade, approximately 50%?

4. Was there a difference in the mean of the assessment scores for the groups related to the particular area of measurement between those students who made up the participants in Free and/or Reduced Lunch

Program grouping of the eleventh grade class, and those students who made up the non-participants in the Free and/or Reduced Lunch Program grouping?

Quantitative Hypothesis

The Total CRT included a locally constructed CRT assessment over taught material and a locally constructed cold-read assessment of reading comprehension/vocabulary proficiency for these students.

A positive correlation of .50 or higher exists between the eleventh grade Quarter Total CRT scores on the locally constructed quarter tests for each Semester 1 and 2, respectively, for Quarter A and Quarter B and the Nebraska Reading Comprehension/vocabulary Proficiency Exam.

A positive correlation exists between the eleventh grade cold-read scores on the locally constructed quarter tests for both Semesters 1 and 2, respectively, for Quarter A and Quarter B in each semester, and the Nebraska Reading Comprehension/vocabulary Proficiency Exam.

Sub-Hypotheses

A hypothesis for each sub-group appears below:

1. A positive correlation exists for those students, approximately 50% of the eleventh grade students, who had completed both quarters of American Literature first semester before the Nebraska Reading Comprehension/vocabulary Proficiency Examination was administered in the spring; even though, these students had been away from formal Language Arts, specifically American Literature, instruction for 10 weeks

or fewer before the state test was administered, scores on the CRT and state reading test and the students' local cold-read scores and the state reading test score.

2. A positive correlation exists for those students, approximately 50% of the eleventh grade students, who had recently entered and were scheduled to complete both quarters of American Literature, but, because of block scheduling, had not had the total amount of time to have completed both Quarter A and Quarter B of Language Arts, specifically American Literature, before the Nebraska Reading Comprehension/vocabulary Proficiency Examination was administered in the spring; thus, these students had completed fewer than 10 weeks of Quarter A of their formal American Literature required course before the state examination was administered, scores on the CRT and students' scores on the state reading test, and the students' local cold-read scores and the state reading test score.

3. A statistically significant difference exists between the mean scores of the male eleventh grade students assessed by the Nebraska Reading Comprehension/vocabulary Proficiency Exam and the mean scores of the female eleventh grade students assessed the Nebraska Reading Comprehension/vocabulary Proficiency Examination.

4. A statistically significant difference exists between the mean scores on the Nebraska Reading Comprehension/vocabulary Proficiency Examination for those eleventh grade students who were participants in

the Free and/or Reduced Lunch Program and those eleventh grade students who were non-participants in the Free and/or Reduced Lunch Program.

Methodology of Study

The methodology of this study began with collecting the quarter CRT scores for each student in American Literature classes during the 2009-2010 school year. Each American Literature teacher, five (5) first semester and six (6) second semester, was asked to give a copy of the Pearson Benchmark Individual Student Response printout to the researcher for Semester 1, Quarters A and Quarter B, and then for Semester 2, Quarters A and Quarter B; thus, 22 total score reports were collected, each labeled with teacher name, student name, date of test, semester and quarter information. These teachers printed off and had in their possession the individual responses each student had made to each question on the test. After receiving IRB approval, the researcher began the disaggregation of information.

The disaggregation of information meant the researcher had to count the number of correct responses for each eleventh grade student on each quarter exam and record them by CRT responses and by local cold-read responses. For example, a semester CRT examination might have 55 questions on the test; the first 35 questions were over the taught material (CRT) and the last 20 questions were over the cold-read materials. Scores were marked, i.e. 29/35 and 17/20, respectively. The researcher could then read the total score of the student

provided by the printout and double check that the two numerators added to that total.

After the researcher had completed the disaggregation of student responses, the researcher then entered the student scores into the SPSS data sheets. The researcher set an order of teachers that was maintained and used each quarter; this assisted the accuracy of data entry process. Then the researcher assigned the students' names consecutive numbers that matched a data line number that was used to hold the information. For instance, the first student in the first teacher's class for Semester 1, Quarter A was assigned the number 1, the second student in that same teacher's class for Semester 1, Quarter A was assigned 2, etc. This process continued for each teacher, each quarter of each semester. The numbers were placed under the student name on the original individual response score sheets. This process then always assured the researcher that correctly labeled information was available for each set of data and was used to check data entry accuracy or to correct incorrect data entry.

All students' information was entered. Some students' information was eliminated later, but initially all student scores for each semester's Quarter A and Quarter B CRT and cold-read tests were entered into the data sheets.

Data columns were headed in the following manner: Semester 1 Quarter A, taught material scores under the header (s1qalit). Semester 1 Quarter A cold-scores were entered under the header (s1qaread). Semester 1 Quarter A total local CRT scores were entered under the header (s1qaTot). Semester 1 Quarter

B CRT scores were entered under the header (s1qblit). Semester 1 Quarter B Cold-scores were entered under the header (s1qbread). Semester 1 Quarter B total local CRT scores were entered under the header (s1qbTot). Semester 2 had the same style of headers, but the semester number was changed, i.e. (s2qaTot), etc.

The researcher then recorded the gender of each student under the heading (sex). The males received the number 1 and the females received the number 2.

The researcher requested a listing of students that participated in the Free and/or Reduced Lunch Program from the school district Executive Curriculum Director. With that information, the researcher entered in the column headed (lunch) the number 1 for participants and the number 2 for non-participants.

The researcher could check the arithmetic score totals for each student by adding the lit and cold-read score for each quarter and checking it against the total score for that quarter. If something did not match, the researcher could check the data against the original sheets with the data entry line number entered by each student name, respectively.

The last column had the header (stateread), and in this column the Nebraska Reading Comprehension/vocabulary Proficiency Examination score for each student was entered.

The research labeled the list of state scores that were produced with the student names with the data-entry line number so that the correct information for

that student's scores could be placed under the correct header and on the corresponding data entry line.

The researcher labeled the list of participants in the Free and/or Reduced Lunch Program with the corresponding student data entry line number so information could be placed under the correct header on the correct data line.

After all data was entered, checked, and rechecked for accuracy, blank data lines were eliminated, and students with incomplete data were eliminated. This changed data entry line numbers and further guaranteed anonymity of student data.

Definition of Terms

The following operational terms were used throughout this study.

AYP: Adequate Yearly Progress is the prescribed progress a qualifying group of 30 students identified to measure was to make to meet the prescriptions for that category or grouping set out by the 2001 ESEA.

Class A school district: In the Nebraska, a Class A school district is one that has a minimum of 450 males, and/or a minimum of 450 females enrolled in grades 9 through 12. In addition, a Class A school can be determined by ranking the 28 schools with the largest enrollments from largest to smallest 1-28 and they are listed as Class A. A school that had always been classified as a Class A school may petition to remain in that Class A designation if enrollment numbers fall below the required minimums or the high school enrollment is not in the 1-28 ranking of largest enrollments. (NSAA Homepage).

Cold Read: In this instance, a cold-read test means the material is deemed at appropriate grade level for reading and comprehension assessment, and it has not been taught to the students before they are tested on it. Each eleventh grade student is to read the same pre-selected readings and answer the same prescribed questions designed to assess the level of reading comprehension and vocabulary proficiency of the students over the pre-selected material (Pennington, 2009).

CRT: CRT is an abbreviation for criterion-referenced test. This type of test was designed to measure what students know from material that had been taught to them. No prescribed distribution of scores was expected. If all students scored 100% on the test, it was an acceptable outcome and within the design of the test. A CRT was not intended to rank students; rather it was intended to measure the student learning of taught material (Bond, 1996).

2001 ESEA: The 2001 Elementary and Secondary Education Act was sometimes called the “No Child Left Behind Act” (NCLB). It established minimum levels of achievement for qualifying public school students. If these levels of achievement, called AYP, adequate yearly progress, were not met after the grace period, then monetary penalties were administered to the schools or school districts.

The Nebraska State Reading Comprehension/vocabulary Proficiency Examination: This test was administered on a statewide basis for all Nebraska eleventh grade students for the first time in the school year 2009-2010. It was designed to test eleventh grade students’ reading comprehension/vocabulary

proficiency in regard to grade level expectations. The test covered cold-read materials, or materials the students had not been taught.

NRT: NRT is an abbreviation for Norm-Referenced Tests. This type of test was designed to rank students from the 1st percentile through the 99th percentile in knowledge of the testing material. Several common NRT exams are the American College Test (ACT), the Iowa Test of Basic Skill (ITBS), and the Scholastic Aptitude Test (SAT) which all rank student achievement for general knowledge on a national level (Bond, 1996).

Qualifying schools: Within the 2001 ESEA regulations, a public school district is held accountable for student learning if the school had at least 30 students in a measurement category. Some of these categories include ethnicity, race, gender, special services, or grade level groupings; i.e. a public school or school district with 29 fourth graders but 30 fifth graders had to show AYP measurements and progress for the fifth grade students in whatever categories apply to the school district, but it did not have to demonstrate AYP for the fourth grade students. This definition, when practically applied, resulted in smaller school districts, as well as, all non-public schools being exempt from the measurement—even if the non-qualifying school had received some federal funding.

Quarter Exams: The particular senior high school used in the study is on a 4 X 4 block schedule. Students may take up to four classes a day for each of four quarters; each class period runs 88 minutes. At the end of one quarter, each student has had the opportunity to complete enough class time and has

been exposed to enough content to qualify for a semesters' worth of credit if requirements for grading were met. American Literature is a full year course, so students are required to earn credit in two quarters—Quarter A and Quarter B. Eleventh-grade Language Arts, specifically, American Literature, is offered two consecutive quarters each semester; thus, the course is equivalent to a full year. Students earn a full year's credit in two quarters: A and B.

STARS: School-based, Teacher-led, Assessment and Reporting System is the Nebraska's response to the 2001 ESEA requirement for accountability through assessment. Nebraska school districts devised their own assessments for their students; thus, each district demonstrated that students met the state standards according to the district-established measurement procedure. Nebraska was the only state in the United States not to have a statewide test for its students in qualifying districts. Nebraska established the Six Quality Criteria for districts to use in making individual district assessments. The local school district in this study used these criteria as the basis for the locally created CRT examinations for Language Arts 9-12.

Background and Rationale

Only the statewide writing test is administered to and measured for all Nebraska public school students; that assessment is prescribed for students in grades 4, 8, and 11 in public schools.

Before this research project, the school district had conducted considerable study, and much money had been and was invested regarding CRT construction and establishment of the CRT statistical validity and reliability. The

local district in this study had committed to locally created CRT examinations for more than seven years in response to the STARS program (Roschewski, P., Isernhagen, J., & Dappen, L., 2006) implemented by the state of Nebraska in response to the 2001 ESEA.

The school year 2009-2010 offered this local school district an opportunity to measure its eleventh grade Language Arts students, specifically American Literature, through the district-devised CRT examinations and then to compare those student scores with these same students' scores on a Nebraska prepared reading test. Until 2010, a Nebraska state test of any kind, given to every eleventh grade Nebraska public school student, had not been administered except in writing. In the spring of 2010, the state of Nebraska required its first statewide reading test, and that was the Nebraska Reading Comprehension/vocabulary Exam, which measures the reading comprehension and vocabulary proficiency of public school, eleventh grade students. This testing afforded the local school district its first opportunity to measure its students against a state "universal standard," as well as to measure its students' scores against other districts within the state to determine if the local district is meeting its own goal for its students. This measure of the relationship between the CRT scores, the cold-read scores and the state reading scores allowed the local district to see what its high school students' reading scores were. The testing offered, as well, an opportunity for the school district to measure its eleventh grade Language Arts, specifically American Literature, tests' results with

the Nebraska test results and discover the relationship between its own measuring tools and the statewide measuring tool.

Delimitations

The Portable Dissertation by Bryant (2004) defined delimitations as reasons that prevent a researcher from stating that the findings from a study always apply to the findings in another person's study or in other situations. The delimitations for this study are the following:

1. The study was conducted in a particular school district with fewer than 300 (278) eleventh grade students rather than a larger study including a larger number of students and school districts.
2. The study is made up of scores from students whose gender numbers were approximately 50 percent males and 50 percent females; an almost even division of gender in a high school grade level population is unusual.
3. The high school offered its course work through a 4 X 4 block schedule, where each student can take a maximum of four classes a day for a quarter. At the end of a quarter, a student has had the opportunity to complete the equivalent of a semester at a traditionally scheduled high school. American Literature, a required eleventh grade Language Arts course, requires a student to take two quarters, one Quarter A and one Quarter B, to meet the course requirement of one full year. This means that approximately half the students in this school district will not be receiving formal instruction in Language Arts, particularly American

Literature, when the state test is given, unlike the majority of eleventh grade students in the state that are in traditional class schedules.

4. The state reading score used in the correlation is based on the first year administration of the examination.

5. The local cold-read test had not been statically tested for reliability but had received test committee consensus for face validity and reliability.

6. The local test's first administration of the reading comprehension test was through a pilot test in the October of 2009, and the second administration of a reading comprehension test through the local pilot test was in March of 2010. Each test was administered to approximately 50 percent of the eleventh-grade, Language Arts, specifically American Literature, students at this local high school.

7. The state cold-read test reliability had not been released when this test was administered. This was the first year for the test. A statewide pilot test was offered in May of 2009, but no Nebraska public school was required to participate, and those schools that did participate did not get results from the state to know then how well their students did.

8. The local school district, in compliance with the state learning objective, had Latin and Greek root vocabulary study formally installed into the curriculum for 2009-2010 sophomores, but the 2009-2010 eleventh grade students did not have this exposure or requirement the previous year.

9. Because of student absence during testing or student withdrawal from school, some eleventh grade students did not take all three examinations:

Quarter A CRT, Quarter B CRT, respectively, and the state reading examination.

10. Because of district and/or state requirements for American Literature credit for graduation from high school, some eleventh grade students in the American Literature classes were exposed to the material and the Quarterly CRT examinations more than once.

11. Because of the district and/or state requirement for American Literature credit for graduation, some students in the classes were not classified as eleventh grade students.

12. The local school district had a fluid student population; students enrolled and withdrew from school throughout each quarter. Because of this fluidity and alternative education opportunities within the district, some students took the CRT examinations after the common date for the eleventh grade students. Also some students had yet to take the CRT examinations when results were released, and some students took the state reading test but had no corresponding CRT scores to accompany it. None of these students' scores were included in the study.

Limitations

The Portable Dissertation (2004) by Bryant defined limitations as reasons that made the study repeatable by another researcher even if years have gone by between the studies. The limitations of the study follow below:

The 4X4 block schedule limited the number of people who could receive an entire semester's (two consecutive quarters) instruction in Language

Arts, specifically American Literature, before the Nebraska Reading Comprehension/vocabulary Proficiency Examination.

Purpose of the Study

The purpose of this retrospective, correlational study was to determine whether the locally designed, state standards-based, American literature criterion-referenced testing was a reliable method of assessing student achievement of Nebraska state standards in eleventh-grade Language Arts, specifically American literature, in total and reading comprehension and vocabulary proficiency.

Assumptions

1. The eleventh grade students in the study were representative of the students in the Language Arts classes, specifically American literature, at this local high school.
2. The textbook used for daily teaching of these students was appropriate for eleventh grade, American literature students.
3. The faculty that taught the eleventh grade students was certificated and qualified to teach Language Arts to this grade level.
4. The locally constructed test had content validity as determined by the curriculum committee through consensus.
5. The locally constructed test measured the identified state standards at an acceptable level of sufficiency—determined by the curriculum committee. That level of sufficiency was a minimum five (5) questions over each measured standard, and the questions over a standard must

appear in multiples of 5. So, 5, 10, 15, etc., questions over a standard must be on the test.

Summary

Chapter 1 described the background and rationale for the study, the purpose of the study, the research question, sub-questions, the quantitative hypothesis, the sub-hypothesis, methodology, limitations of the study, the delimitations, assumptions of the study, and definition of terms used throughout the research study.

Chapter II: Review of Literature

Introduction

In this review of literature, readings pertinent to the study as well as a review of the affects of the 2001 Elementary and Secondary Education Act (ESEA) are considered. This United States Education Act, in particular, was most responsible for the resurgence of interest in a demonstration of successful teaching and student learning in U.S. public schools. The whole accountability philosophy exerted much influence in U.S. education and had its re-emphasis beginning with the passage of the 2001 ESEA (Ferrandino, 2006).

After the initial discussion and explanation, the chapter discussion focuses on the Nebraska Department of Education's response to this 2001 ESEA. Nebraska's response was the STARS program--Standards-based, Teacher-driven, Assessment, and Reporting System--a program unique to Nebraska (Dappen & Isernhagen, 2005).

Finally, the chapter focuses the discussion on the fact that because of the 2001 ESEA's requirement for nation-wide measurable accountability in public schools, research studies evolved throughout the nation that were designed to establish the validity and the reliability of measurement. Many districts and individual educators have used these data to establish the credibility of methodology used to teach U.S. public school children.

2001 ESEA Historical Perspective

The 2001 ESEA sent U.S. public school educators into a busy and necessary response because educators and districts were expected to show

measured student progress or achievement. The educators felt the urgency to show satisfactory student progress. These same educators knew a threat existed for their schools or the entire school district to have federal or state funding taken away or some funding reduced. The educators in their respective states rapidly developed student assessments that addressed, met, or exceeded the expectations set out in this 2001 ESEA; specifically, the Adequate Yearly Progress (AYP). The AYP's function was designed to be a concrete manner of assessment to hold a school accountable for student progress.

If a qualifying public school district had any group of 30 or more students in a prescribed, measureable group, such as grade level, special education, ethnic group, subject area, etc., it was considered a viable accountability grouping. The viable group was then measured for AYP and the results were recorded, and these results became integral determiners in the funding formula. If the viable group did not meet the required progress, the school or district was labeled failing, and, under certain conditions, federal and/or state money for that grouping was threatened to be reduced from the school or district budget. Or, if this failure happened for a prescribed period of time, the funding was permanently eliminated. If more than one grouping was determined to fail to meet the expectation, all federal and state funding was threatened to be eliminated for the next year. Some facets of this act that concerned educators are the following: If, on the day of student testing for AYP, a student or students were absent, that/those absence(s) counted against the school district's AYP score, and, if that student's absence made that group fall below the number thirty

or 30 students minus 2% that were tested, the group was listed as not having made its AYP; thus, that school district's funding was jeopardized.

Small public districts (districts with fewer than 30 students in every group) were excluded from AYP, or more precisely, if a public school district did not have 30 members in a prescribed or defined group, that district was not held accountable for AYP for that grouping. Educators asked questions: If this 2001 ESEA was created to guarantee that good, minimum levels of learning opportunities existed for U.S. school children, what difference did the size of the school or the size of the group make?

The prescribed progress was another important concern. Schools were given the percentage or number of students who had to meet the minimum score, and the date this achievement had to occur. For example, if a school district initially had 60% of its fourth grade students scoring at the 80th percentile on the yearly reading achievement test, and, if the 2001 ESEA requirements stated that those achievement scores had to show that 95% of the students had to meet the 80th percentile by end of the next school year, even if 94% of the students met the requirement, the school was listed as failing its AYP for that grade on that measurement. Financial penalties were then a possibility for the school district. However, if one or more students moved into a different school in a district in one grade making 30 or more students--thus the AYP group immediately before the testing time frame, those students' results were included in the group AYP. Students had to be enrolled in the district from the 4th week of

September until the test date to be included in the results, but they could move from one school to another and be counted in the test school's AYP (NDE).

Another concern for educators was that The 2001 ESEA demanded that student achievement improvement was made each year. So, if a school district had a viable group achievement measurement at the 99% percentile, and, if that same group did not raise its achievement during the following school year, the group was determined to fail in AYP.

Perhaps the largest concern for public school educators was the fact that only U.S. public schools with viable group sizes of 30 or more were held accountable for student progress. Were not all U.S. schools, whether private or parochial and large or small, meant to be accountable in education for the good of the children?

From the beginning, this 2001 ESEA was nicknamed, by the then United States Department of Education Secretary Margaret Spelling, as the No Child Left Behind Act. The intent of the act was that each state confidently guaranteed that each child within its jurisdiction received a minimum, good public school education. Because built-in deadlines were included within the 2001 ESEA regarding student progress, schools were asked to measure and record the prescribed student progress, and to be aware of financial penalties if the students did not show required, appropriate progress by those deadlines. The new and stringent expectations of measurement and the reporting of achievement caused organizational costs and time concerns.

In the education community the NCLB Act was credited as the reason for the instant, intense, high priority efforts placed on student assessment as the measure of school accountability. Often these tests or assessments were labeled "High Stakes Testing" (Sloane & Kelly, 2003). This label was applied because federal aid to schools was tied to the results of the tests, or sometimes the label was used because the test results were published in papers. Although the pressure to measure progress, to know the level of student success, and to meet or exceed the level of success was overwhelming after 2001, accountability was not new to the education process.

Education Historical Accountability

From the Puritan Era to the 21st Century, accountability for learning in a group setting was always an element of public teaching and student learning. The importance of the accountability factor vacillated in this time frame from subtlety to prominence, but its presence was always felt in the education field. As Budnik (2001) discussed, during the Puritan Era, The Massachusetts Bay Colony expected each of its member towns to teach each child to read so each Puritan child, and eventually each Puritan, was able to read the Bible; and, like the NCLB Act, accountability for no progress took the form of a monetary fine. Accountability is not just a one-time event in history; Haladyna, Haas, and Allison (1998) highlighted ongoing examples of assessment and accountability, and they described these as a part of the U.S. system of public education.

The impetus for standardized tests emerged in the 1800s and continued. . . . The first documented achievement tests were administered in the

> period 1840 to 1875, when American educators changed their focus from educating the elite to educating the masses (p. 2).

As Haladyna, et. al., (1998), further pointed out and referenced other education historians in his writing, specifically (Cremin, *The Transformation of School: Progressivism in American Education*, 1876-1957, New York: Vintage Books):

> . . . the earliest tests were intended for individual evaluation, but test results were inappropriately used to compare schools and children without regard for non-school influences. As millions of immigrants came to the United States in the 19th century, the standardized test became a way to ensure that all children were receiving the same standard of education. In fact, however, test results were often used to emphasize the need for school reform (p. 2).

Accountability for teaching students continued on throughout history. Sometimes the accountability was measured by achievement tests or assessments; sometimes accountability was measured by intelligence tests, which then shifted the accountability to the individual student rather than the school district, or the curriculum, or the teacher, but it was measured accountability, nevertheless. Since accountability became associated with the U.S. education system in its many forms and for a such a length of time, it seems obvious that the NCLB Act did not create the accountability mode in U.S. Education, but it certainly changed the mood toward accountability, and the 2001 ESEA placed renewed emphasis upon accountability.

For years federal or state aid was distributed to districts based on rules and goals, but a prescribed AYP was not in the criteria, and a prescribed level of success was not imposed upon a district for each category of student groupings as it was after the NCLB Act was implemented. The 2001 ESEA made funding incumbent upon the prescribed and measured successful student accomplishments. School districts became increasingly, if not exclusively, interested in the development of tests or other assessments that demonstrated that the district, the school, and/or the student had met the prescribed mark of accomplishment.

Assessment Selection Difficulty

The assessment discussion for educators in the U.S. became centered upon what was the better test to give? Was a norm-referenced test (NRT) a better test than the criterion-referenced test (CRT) because it was to be used nationally to see where students all across the nation ranked against each other, or, should the student assessment be conducted through criterion-referenced tests which were designed to see how much the student knew of the material that was specifically taught (Bond, 1996; Huitt, 1999)?

For the NCLB Act's testing, states and individual school districts were allowed to decide which test type each believed was the better form of assessment for its students. Of course, for the education community, the assessment process still was not decided. The whole process of assessment and test selection became more complicated, and groups formed into camps of support for one type of assessment versus another.

The question still needed answering, whether NRT or CRT was better. If NRT were the method, which test was better: CAT, ITBS, or ACT? Or, if CRT were the format selected, was it to be a district-made test, or was it to be a test purchased from an educational testing service? Or, was the CRT testing to be done through portfolio assessment, and, if so, was this method really better for student assessment? What about performance-based assessment, or authentic assessment? It needed to be determined, which was the better assessment format, and why it was better. "When confronted with . . . NCLB Act and . . . AYP requirements, every state but Nebraska decided to use norm-referenced or state selected CRTs—'high-stakes measures'" (Dappen & Isernhagen, 2005, p. 147). Most states examined what NRTs or CRTs were available for purchase; some made a selection and bought a test that best suited the state's needs, and each state assessed the entire state's body of qualifying public school students' learning with these tests. Some states formed committees and created a statewide test to be used by every public school within its borders to assess its student progress.

Theoretical Historical Basis

Nebraska decided to leave the question of which test to use for student assessment to each school district to decide rather than a state-mandated test for all districts. Nebraska met the "letter and spirit" of the U.S. Constitution and the education law when it allowed each district in the state to be responsible for education and assessment of its students. Instead of following the pattern established by the other states, Nebraska formed a standards committee to

develop state standards that were used by each public school district in the state when that district was drafting its assessments (Roschewski, Isernhagen, & Dappen, (2006). A school district could develop its own local standards that were equal to the state standards. This program was called STARS: School-based, Teacher-driven, Assessment, and Reporting System. The school districts were required to devise a portfolio of instruction that was submitted to the state for evaluation and acceptance. The portfolio showed what the district was teaching and how this portfolio met or exceeded the standards. Because Nebraska chose a unique method of selecting school or district accountability measures, Nebraska still had other issues to consider.

First, Nebraska chose Language Arts and math skills as the first subjects tested on a statewide basis. For Language Arts, the state selected grades within school levels for statewide testing: elementary, middle school, and high school. Teachers incorporated the local school district's values and commitments and determined the local standards that were to be met in conjunction with the state standards, and that students were to be measured against the standards by predetermined quality benchmarks. These benchmarks or measurements were then submitted to the state committee designated by the Nebraska Department of Education. Some districts used portfolio assessment, and some districts used CRT examinations, which were either purchased or district made.

Each district was held accountable and required to self-report student progress. In addition, portfolios created by school districts were offered as a record of what was to be taught in that district and, on a random basis, these

portfolios were sent to a completely independent committee that was outside the state for purposes of evaluating the portfolio's quality and effectiveness.

Nebraska did use one statewide test, and that was the state writing test for students in grades 4, 8, and 11. Each tested grade level wrote a required essay type for the test: narrative, descriptive, and persuasive, respectively. Similar to the portfolio quality control, a random sampling of writings from all three levels of the state assessment were sent to an outside, independent committee to be reassessed after the state-wide writing assessment and evaluation occurred; this outside assessment worked as a quality-control safeguard for the state, the teachers, and the students. The quality control efforts confirmed that the students were producing at a universally recognizable, minimum level of achievement, and that the teachers who evaluated were as effective as the teachers of other states in their evaluation skills.

Nebraska continued to assess in this manner after 2001; but, for the school year 2009-2010, Nebraska added a required reading comprehension/ vocabulary test for all Nebraska public school students in the eleventh grade. This, too, met the guidelines of Nebraska standards, and the test was subjected to the outside evaluative, quality-control committee.

Questions Not Answered

Even after assessment selections were made by each state, those state's educators were still in search of answers for hypothetical but legitimate questions. These educators had many concerns about the impact of group assessment (McGehee, J., & Griffith, L, 2001) and they continued with the

implementation of the NCLB Act on their communities, districts, or classrooms (Gallagher, C., 2004). Articles were written that emphasized basic educational philosophy that was ignored. Some authors questioned “quick fixes” offered by non-educators, and others offered staunch resistance to change to the traditional process of school. Ferrandino (2006) summed up a concern that was pervasive in the education community:

> We all know students who just want to figure out what it takes to make the grade. . . . “What do you want to get out of this class?” and their answer is, “An A, or . . . maybe just a passing mark. . . .” As the NCLB Act had increased the pressure on school across the country to meet the multi-pronged definition of “adequate yearly progress,” the single-minded emphasis on making the grade was a temptation for school leaders as well (p. 1).

Teachers were very concerned that the intent of 2001 ESEA was to limit student learning; many teachers believed that the refocus of all student learning toward the achievement of a minimum education was the end of quality education, and clearly the demise of academic freedom, individual student excellence, teacher and student creativity, and in-class spontaneity in the public schools.

When district assessments were given, the published scores--state-by-state, district-by-district, and school-by-school--caused further concern. Because the scores were often compared to scores from other schools or other districts, the published scores caused consternation. Even states that gave different

assessments were compared to each other. Judgments were made regarding the quality of education that students in that state, that district, or that school received. Experts (Haladyna, Haas, & Allison, 1998) noted that these published pieces caused celebration, which was passed on to the superintendent; or, blame, which was then passed on to any and all people in positions below the superintendent.

Attitudinal Evolution

After it became apparent that the U.S. Congress was not going to modify the NCLB Act to give schools more time to develop assessment processes or more leeway not to assess students, educators wrote articles that described preparation for mandated testing without the abandonment of education values. Some educators questioned the effects upon the professional staff now that universal testing was used as input in teacher evaluations (Haladyna, Haas, & Allison, p. 6). And, educators often wrote that the scores from assessment needed to be used correctly; instead of using the scores as fodder for the "our school was better than your school competitions" that historically plagued U.S. education. "The misinterpretation and blatant misuse of test scores is pervasive" (Haladyna, Haas, & Allison, 1998, p. 6).

How the tools were to be used caused great concern. And, concern was shown regarding the results provided by the tools to evaluate or assess. As Robert Schultz (2000) noted, the same tool was used for either evaluation or assessment. In the case of a rubric, if it were to be used in a summative assessment, then a grade was assigned based on successful achievement of the

sections of the rubric. If the same rubric were used as a formative assessment, then it was used to identify areas of strength and weakness and highlighted these areas for use in student correction or remediation. It was the intended use and the point in the student learning process that was the determining factor of the use and value of a tool. Eventually, within the education community, consensus occurred for many if not all types of assessment tools, which eased some angst about measurement and allowed the education community to direct its focus to the teaching of students in the best and most efficient way and to measuring student learning.

Research Leads to Solution

Because of the concern in the education world, research was conceived and conducted to help discover if a better method of instruction, assessment, or student grouping (Espin, Shin, & Bush, 2005, p. 3) made a positive difference in student learning, or knowledge retention, and/or assessment performance.

But, as time progressed and educators adjusted to the NCLB Act's expectations, educators began to change their attitudes towards assessment. The educators were not just satisfied with the aim of student achievement as meeting minimum AYP. Once teachers adjusted to the impact of the regulations, they began to devise the best method of assessment of student learning and achievement. School districts and teachers committed to educating their students to the highest level, rather than just settling for the minimum level of achievement. In some cases, the state's new minimum level of achievement was aimed at a much higher level of achievement than had been attained before the

2001 ESEA, but the minimum level of achievement for the 2001 ESEA was still not the final desired level of achievement for the education community. Educators admitted, “Teaching begins with assessment” (DiRanna, Osmundson, Topps, & Gearhart, 2008, p. 23). The teachers began to examine where they wanted to be at the end of the unit or year, and they designed teaching and assessment to allow themselves and the students to achieve this.

School districts, other than those in Nebraska, found that sometimes the test that was used as an assessment tool was not compatible with the educational approach and/or the subject matter selected for the district’s students (Vu, 2007, p. 1). Educators discovered that sometimes the test did not benefit the student group being assessed (Frisbie, Miranda, & Baker, 1993). More questions came forward. What subjects in school should be assessed first? What grade levels should be the first to be measured? The education community asked itself this question: “How do we know that we measured what we intended” (Ferrandino, 2006, p. 2)? Other questions that probably were asked include the following: What education or local principles or values were lost in the teaching process when education focused on assessment? What, if anything, did students gain in the change of focus? What was the best way to assess student learning?

Educators engaged in research both qualitative and quantitative (Dappen & Isernhagen, 2005) because they sought understanding of the purpose of assessment, and they sought to test validity of the assessments. Espin, Scierka, Skare & Halverson (1999) did research on secondary writing programs.

These researchers used curriculum-based measurements to see if the students made progress in writing and to see if these assessments predicted a student's writing proficiency on Oregon's state writing exam. Espin, Scierka, Skare, & Halverson (1999), conducted research that espoused the philosophy of another set of researchers, Downing and Haladyna (1997), that any research required "using a careful and systematic approach . . . and careful documentation, record keeping, and a method of systematic, routine reporting of this documentation" (p. 63). As Espin, Scierka, Skare, and Halverson (1999) discovered, if the students made progress, then no change was made in methodology or materials, and the assessment was considered valid.

If, on the other hand, the students made no measured progress, then the assessment was considered to point out invalid teaching methodology and/or inappropriate materials. One or both of these conclusions caused appropriate change to match more closely the standard or methodology needed. This was a common approach in the application of the research. After finding this method of research practical and helpful, these researchers went on to use some of the methods used for writing assessment validity and proficiency predictability in the elementary school setting for the high school setting, too. Here, too, they found validity. The test measured what it was supposed to measure, but the researchers found that a combination of variables was needed to predict writing proficiency in high school students.

In addition to the validity of assessments, school personnel were interested in the reliability of the assessments. Educators wanted to know that

their students achieved similar results each time these students took the test or similar tests. Often the researchers ran a correlation between the pre-test and the posttest, or between the raters of writing exams, or a particular grade level in consecutive years. If the participants' scores showed a positive correlation of above .50, the test was considered reliable. If the correlations approached .80 or better, a high reliability was concluded (Espin, Scierka, Skare, & Halverson, 1999).

Student writing was another form of reliability of assessment that state education departments and school districts wanted. Writing was generally considered to be personal, style was individual, and rating was thought to be subjective, so, clearly, writing assessments needed the reliability established in such a way that it produced a level of confidence and ensured the assessment's predictability to identify successful assessment and proficient writers. A research study conducted by Espin, Wallace, Campbell, Lembke, Long, and Ticha, (2008) which expected to predict the success of high school students on state standards tests was conducted. Students wrote for 10 minutes in response to a narrative-writing prompt. Students' progress was marked at three, five, and seven minutes. The assessors used evaluation criteria that included word count, correct spelling, beginning and ending of sentences, and correlated the information with the students' performance on the Minnesota Comprehensive Assessments. "This research showed a correlation range of .64 to .85, and the strongest coefficients occurred with the 10-minute writing timing. This type of empirical research made a confident test administrator and a credible test" (p. 6).

Inter-rater reliability research was conducted (Herman, Gearhart, & Baker, 1993) and established for writing assessment. This type of research was instrumental in furthering the development of credibility in the assessment process.

Other types of research were conducted in school districts to establish trustworthy assessment tools. Meeting or exceeding the standards of the NCLB Act was always the goal, but the variety of research suggested that school districts and personnel had a creative, altruistic interest. Each research project conducted was in regard to assessment in terms of validity, reliability, alternative nature, portfolio status, or for the purpose of investigating performance-based assessment, and authentic assessment. All research had a common purpose to see if it provided more information that allowed finer and finer honing of education methodology or assessment design so the measurement was more informative and accurate.

A study by Fien, Baker, Smolkowski, Smith, Mercier, Kame'enui, Beck, and Thomas (2008) covered reading fluency and its evaluation for second graders: both English learners and native English speakers In this study the researchers tested a theory about the predictability of reading proficiency through the use of nonsense word fluency and found that a high correlation occurred between the state test and the national test for predictability of fluency for these students. This type of information made achievement expectations more likely and provided some classroom teachers with support for their beliefs that all students were able to learn to read.

Another detailed study conducted by Frisbie, Miranda, and Baker (1993) was designed to evaluate elementary textbook tests to see if the tests were quality tests. The research concluded that most of the questions were, according to Bloom's Taxonomy, at the knowledge level. These researchers recommended that the teachers not use the tests that accompanied the texts. This information reported to the textbook company affected change in assessment design by the company. The information assisted learning and assessment for that year and other years because it enhanced the level of testing. The feedback of the textbook tests to the textbook companies encouraged teachers to carefully evaluate the textbook tests.

Research was designed for a variety of aspects of teaching and evaluation. One researcher, Foote, (2007) designed a study that was conducted in order for schools to ascertain that the teaching was aligned with standards and in effect helped educators by "Keeping Accountability Systems Accountable" (p. 1). Accountability and watchfulness maintained the integrity of the school system. The aim was to validate the success of the program and validate the goal attainment for each student involved. As of 2009, more than 29 of the states required graduation or exit tests for its students, which added to the credibility of the state's education, the school district's credibility, the teachers' effectiveness, the students' achievement, and the policy makers' judgment (Zhang, 2009). To maintain credibility of the tests, they must be regularly monitored for standards alignment.

Research has been conducted that considers whether young children should be assessed (Bordignon & Lam, 2004). The research varied from the traditional research that tried to discover whether there was a benefit to testing individual students in novel, neutral, or reward situations (Christ & Schanding, 2007), to other forms of education research, which included whole school evaluation (Crowley & Hauser, 2007). And, sections of subjects were researched, as well. Research was conducted to see if vocabulary matching indicated learning in social studies (Espin, Shin, & Busch, 2005). A consistent and constant pursuit of quality information to augment the collective body of knowledge continued to be a worthwhile focus. The increased body of knowledge suggested that quality teachers demonstrated integrity in education through valid, reliable assessment and confirmed teachers were a valuable asset to the education process. The research provided evidence that educators had enriched their students' learning opportunities.

Once accountability was accepted as a positive facet of teaching rather than invasive, research was produced in many areas of education. Valid, reliable methodology for CRT evaluation became critical. Varied forms of assessment, not limited to just multiple-choice tests (Ogawa, 2003), emerged from sound research. Portfolio research and appropriate, authentic assessment research (Rivera, 2005), learning styles research, teaching styles research, open-ended answers tests research, writing assessment research--all became objects of focus, which benefited the education field. Literally every public school district in every U.S. state and some U.S. territories was involved in adopting accountability

in its student and faculty community. Each research project refined the knowledge and set a more accurate measure of progress (Hintze & Christ, 2004) that these schools or school districts used with confidence.

Through historical reference and commitment to quality education by dedicated education professionals, student assessment for student learning accountability in education changed the accountability journey from a dreaded, required duty to a highly regarded facet in the development of best practices of education the United States.

Summary

Chapter II discussed the emphasis on accountability that the U.S. Public School System has faced since the 2001 ESEA was enacted in U.S. Public School Education, the diversity of educational research that came forth from this refocus on accountability, and the change in attitude of the U.S. educators after they became accustomed to the expectations of the 2001 ESEA.

Chapter III: Methods and Procedures

Introduction

Chapter III outlines the methodology of the study and the study's design, purpose, examination choices, and details of the examinations. These details include how the subjects were selected for the study; the grade level they occupied during the study; what, if anything, affected their full course of study in the Language Arts, specifically American literature; and enrollment status at the time of the state reading test and the results for each quarter CRT examination. The chapter also shows the tests' correlational relationships between the locally designed quarter exams and the Nebraska Reading Comprehension/vocabulary Proficiency Examination. The chapter describes the demographics of the students in the groups and sub-groups and explains the limitations regarding composition of groups as well as the ethical components utilized in the study.

The chapter also explains the procedures and methods, the selection of the instrument for the relationship study, the procedures utilized in the study, the data preparation process for the relationship instrument, the analysis of data, the procedures used to analyze the data, and the process by which the data were interpreted. Part of this study's methodology will consider the limitations of the study.

The chapter was divided into the following sections:

1. Design and purpose of the study including the rationale for this relationship study;
2. Population and confidentiality restrictions;

3. History of the development of the local school district's instruments: the two parts of the local examination.
4. Population and its distribution between semesters;
5. Data analysis.

Ethical Considerations

Student names and identification numbers were needed for the researcher only to verify that the information attributed to each student was correct, but, for the correlations, no student names or student identification numbers were used. Only the scores were correlated.

Each student's data were entered; the first student in the first teacher's class received the number 1 and then scores for each category in the study were entered for that student; the second student in that teacher's class and scores were entered on the number 2 data line and for each category in the study and so forth. Each teacher's class's students were numbered consecutively, so students were assigned data line numbers by their alphabetically arranged names by individual class. Because of this method, no student identities were entered into the data to be examined or divulged in the research.

Study Design

The design of the study was a retrospective, non-experimental, non-randomized, post-test analysis between two examinations to determine whether there was a positive, statistically significant correlation between students' scores on the locally constructed district examination for American literature in total and the cold-read, and those same students' scores on the Nebraska's reading

examination, which assessed reading comprehension and vocabulary proficiency. The data used for this retroactive study were taken from the test results and organized onto a spreadsheet (Appendix A).

Data were arranged on the spreadsheet in the following manner: the first set of data (Table 1) used in a correlation with the state reading scores was the student scores earned on the locally constructed, total criterion-referenced examination which covered the taught material and a cold-read for the school year 2009-2010, Semester 1 Quarter A. A second set of data (Table 1) used in a correlation was the student scores earned just from the locally constructed, cold-read test measuring reading comprehension and vocabulary proficiency 2009-2010, Semester 1 Quarter A. The third set of data (Table 2) used in a correlation was the student scores earned on the locally constructed total criterion-referenced examination over the taught material and a cold-read for the school year 2009-2010, Semester 1 Quarter B. A fourth set of data (Table 2) used in a correlation was the student scores earned just from the locally constructed cold-read test measuring reading comprehension and vocabulary proficiency 2009-2010, Semester 1 Quarter B. A fifth set of data (Table 3) used in a correlation was those scores earned on these same tests by the second semester students Semester 2, Quarter A Total CRT test. A sixth set of data (Table 3) used in a correlation was the student scores earned just from the locally constructed cold-read test measuring reading comprehension and vocabulary proficiency 2009-2010, Semester 2 Quarter A. A seventh set of data (Table 4) used in a correlation was those scores earned on these same tests by the second

semester students Semester 2 Quarter B Total CRT test. An eighth set of data (Table 4) used in a correlation was the student scores earned just from the locally constructed cold-read test measuring reading comprehension and vocabulary proficiency 2009-2010, Semester 2 Quarter B. A ninth set of data (Table 5) established and compared the mean scores from the Nebraska Reading Comprehension/vocabulary Proficiency Exam for each student's gender: male or female. A tenth set of data (Table 6) established and compared the mean scores from the Nebraska Reading Comprehension/vocabulary Proficiency Exam for each participant or non-participation in the Free or Reduced Lunch Program.

A *t*-tes*t* was run using each gender's mean score from these same results on the Nebraska Reading Comprehension/vocabulary Proficiency Exam to see if a statistically significant difference in means existed between these two groupings.

Another *t*-test was run using the means from the scores earned on Nebraska Reading Comprehension/vocabulary Proficiency Exam scores for the students' who participated in the Free and/or Reduced Lunch Program status and scores for those of the non-participating students in the Free and/or Reduced Lunch Program to determine if a statistically significant difference in means existed between these two groupings.

Test Selection

The tests selected for the study were the locally constructed Language Arts, specifically American literature, examination, the CRT section used to

determine the Total Quarter score, the cold-read section, and the Nebraska Reading Comprehensive/vocabulary Proficiency Examination. The rationale for the choice of these examinations was threefold: (1) the state test was in its initial year, and it was the first opportunity for this type of relationship study, and (2) this locally constructed test was in its first year of containing a cold-read test to measure student comprehension and vocabulary proficiency (up until the year 2009-2010 the school district had used locally constructed CRT examinations, but no cold-read section had been designed), and (3) the relationship of socio-economic standing, as determined by participants in the Free and/or Reduced Lunch Program, and the Nebraska state reading examination.

Because the opportunity to examine the local school district's cold-read scores and their relationship with the state test scores presented itself for the first time, and, because the local school district wanted to know the relationship between the state cold-read test with its own Total CRT scores and cold-read scores, it was determined that this study benefitted the district and added to the general body of knowledge.

The research study questions were best answered by finding the relationship of the scores through a correlation of the each semester's quarter test scores, eight correlations in total, and a t-test for significance in gender mean scores and significance in lunch status mean scores on the state cold-read test. The Pearson Two-tailed Correlation was selected for the correlational aspect of the study.

The scores from the entire eleventh grade student body were intended as subjects in the study. These students took Language Arts, specifically American literature, classes through a 4 X 4 block schedule. Approximately half the students took an entire academic year of instruction within the first semester of the 2009-2010 school year. The other, approximately, half of the students took their entire academic year in the second semester of the 2009-2010 school year.

The test scores were used from each semester's Quarter A and Quarter B, and the state reading examination to find the relationship between the eleventh grade students' scores on the locally constructed Total CRT and the locally constructed cold-read examination for each semester's Quarter A and Quarter B.

Only the students who had completed or were enrolled to complete two quarters of the American literature course and were classified officially as eleventh graders took the Nebraska State Reading Test for the academic year 2009-2010.

The correlation tool that was used to demonstrate the relationship was the Pearson Two-tailed correlation instrument because it is a more stringent test and less likely to wrongly reject the null hypothesis.

The *t*-test was used to determine the statistical significance between the mean scores for the students according to the appropriate groups of gender and lunch status.

Student's scores from each semester were added together to create a population of 278 students who took both the locally constructed examination and the Nebraska Reading Comprehension/vocabulary Proficiency Examination.

The local school district required its eleventh grade Language Arts, specifically American literature, students to take a quarter exam of the locally constructed CRT over material covered in the course and a cold-read test twice per semester, once each quarter. Within the two semesters of the school year, all the students at the eleventh grade level were scheduled to complete both Quarters A and B and be tested.

The Nebraska required all 2009-2010 eleventh grade students in the state to take the Nebraska Reading Comprehension/vocabulary Proficiency Examination once per year within the school days between March 31, 2010, and April 30, 2010.

Students in first semester, eleventh grade Language Arts, specifically American literature, were out of formal classes for fewer than 10 weeks before the state exam. Conversely, those students enrolled in second semester, eleventh grade Language Arts, specifically American literature, classes were in formal classes for fewer than ten (10) weeks before the Nebraska Reading Comprehension/vocabulary Proficiency Exam was administered.

Each student received two scores for the quarter: one for the locally created, criterion-referenced exam and one for the locally created, cold-read examination; these two combined to make a total score for each nine weeks;

plus, students received one score on the Nebraska Reading Comprehension/vocabulary Proficiency Examination.

Study Population

The eleventh grade student population was divided into two approximately equal groups for instruction of eleventh grade Language Arts, specifically American literature. One-half of the population was randomly assigned by the high school registrar through a computer program to take the eleventh grade American Literature course in the first semester of the school year, and the second-half of the population was also randomly assigned by the registrar through a computer program to take the American Literature course in the second semester of the school year, and all students in the study were required to take the locally created quarter tests that corresponded to the course of study and the time the course was taken, and the Nebraska Reading Comprehension/vocabulary proficiency Examination, which was administered only one time per year. For students in a 4 X 4 block schedule, the state test did not necessarily correspond to the semester the student was in formal Language Arts class, nor would the students be at that same point of instruction in the semesters.

Approximately 50% of the population took the American Literature class first semester, and 50% took the course second semester. Approximately 50% of the population was male and 50% was female. Approximately 50% of the population participated in the Free and/or Reduced Lunch Program and approximately 50% of the population were non-participants in the program.

Procedure

Data collection

Collection and analysis began when student scores on the local test were paired with the individual's score on the state test. The student's scores were placed on a spreadsheet under the following categories of Semester 1, Quarter A CRT (s1qalit), Semester 1, Quarter A cold-read (s1qaread), Semester 1 Quarter A Total (s1qatot), Semester 1 Quarter B CRT (s1qblit), Semester 1 Quarter B cold-read (s1qbread), Semester 1 Quarter B Total (s1qbtot). Semester 2 Quarter A CRT (s2qalit), Semester 2 Quarter A cold-read (s2qaread), Semester 2 Quarter A Total (s2qatot), Semester 2 Quarter B CRT (s2qblit), Semester 2 Quarter B cold-read (s2qbread), Semester 2 Total (s2qbtot), State Reading (stateread), Gender (sex), and Free and/or Reduced Lunch (lunch) (Appendix A).

Student scores were accurately recorded and double-checked that each score was accurate for the student. To make certain of the anonymity and accuracy, the student names and corresponding student identification numbers were removed from the study data and replaced with the appropriate spreadsheet line numbers to guarantee anonymity of scores for the purposes of this study. There was no segregation of the students in the correlational study except for the data collection for quarter and semester to guarantee accuracy.

Because eleventh grade American Literature is a required course for high school graduation from the local school district and/or the state, all scores for students, no matter their classification in the education system, were included on the original data collection sheet. However, unless a student had a score for the

Quarter A CRT, the Quarter B CRT, and a score from the Nebraska Reading Comprehension/vocabulary Proficiency Examination, the student information was not used in the study.

Test Rationale

The rationale for testing groups for sub-question analysis was to determine if eleventh grade students who received less than the full course of study, because of second semester registration, had a significantly different result than those who had the full course of study. In addition, an analysis was used to determine if gender or socio-economic status, measured by participation or non-participation in the Free and/or Reduced Lunch Program made a difference.

Content validity of the tests administered to the students was established by the test construction committee, which came to consensus regarding the content in a question by question and standard by standard scrutiny of the locally created test, and a similar process was used by an expert committee for the state exam as well as some purchased questions from a credible testing service.

The locally created tests for reading were created to be similar in format to the state examination. The experts on the locally created test committee were teachers in the school system that taught at least one section of the eleventh grade Language Arts, specifically American literature, course of study. The purpose of creating a similar format was so no need for adjustment to format was required for the students between the local tests and the state test. A member on the committee for the locally created examination was also a member on the

state committee that created the Nebraska Reading Comprehension/vocabulary Examination; consequently, the local committee used similar methods in designing the local test and the construction of it. This methodology was used throughout the entirety of the tests.

The CRT examinations were designed to offer the student four multiple-choice answers for each question; three of the choices for each of the questions were detractors. The students selected an answer and marked a Scantron© sheet with the student selected answer for each question on each of the local exams. For the Nebraska Reading Comprehension/vocabulary Examination, the students selected an answer from multiple-choice options for each question on a computer delivered test. The reliability for each examination was not released as of May 2010.

Data Analysis

After the locally constructed tests and the state test were each administered, the results of the tests were placed in the SPSS software data analysis program for tabulation and analysis. The results were analyzed and then exported to a Word file. The spreadsheet data were used in different correlation combinations. Each correlation set used the Pearson Two-tailed correlation. The students' scores used in the study had to reflect a Quarter A Total CRT score for Semester 1 and 2, and a Quarter B Total CRT score for Semester 1 and 2, a Quarter A cold-read score for Semester 1 and 2, a Quarter B Cold-read score for Semester 1 and 2, and the Nebraska Reading Comprehension/vocabulary Proficiency Examination score. The scores on the

Nebraska Reading Comprehension/vocabulary Proficiency Exam were then used to identify the male and female mean scores on the examination and the mean scores for the Free or Reduced Lunch Program Participants and Non-participants. These mean scores were correlated to see if a statistically significant difference in success on the Nebraska Reading Comprehension/vocabulary Proficiency Examination occurred between genders. The second set of mean scores was correlated to see if a statistically significant difference in success occurred between the participants and non-participants in the Free and/or Reduced Lunch Program.

Summary

This study was designed to be a retrospective, correlational study of student performance scores from the locally constructed CRTs, student performance scores from the locally constructed cold-read test, and student performance on the first ever Nebraska Reading Comprehension/vocabulary Proficiency Exam. Because the senior high had approximately 300 eleventh grade students who must take Language Arts, specifically American literature, over a semester's time, the students were randomly divided into two approximately equal groups: Semester 1 and Semester 2.

The study was to find the relationship among these quarter tests and the Nebraska Reading Comprehension/vocabulary Proficiency Examination through correlations and to reveal whether any significant differences occurred between males and females on the Nebraska Reading Comprehension/vocabulary Proficiency Exam, and to reveal if any significance occurred between Free and/or

Reduced Lunch Programs participants and non-participants on the Nebraska Reading Comprehension/vocabulary Proficiency Exam. Chapter III detailed the methodology, procedures, participants, demographics, and ethical consideration data collection, data analysis and the study timeline of one school year, 2009-2010.

Chapter IV: Results

Introduction

Chapter IV will discuss the findings of this research study. A description of the analysis and the computer program that assisted this study, the data results and tables of correlations and *t*-tests, and, if results were significant at the $p = <.05$ level of confidence, results will be highlighted by asterisks.

Test Selection

As stated in Chapter I, this study examined the relationship between scores earned by the eleventh grade student in Language Arts, specifically American literature, Quarter Total CRT scores, and the quarter cold-read scores, with those results of the Nebraska Reading Comprehension/vocabulary Proficiency Exam through correlations using the Pearson Two-tailed Correlation. The data were placed in the SPSS quantitative software program for analysis.

Pearson Two-tailed Correlation and *t*-test

This Pearson Two-tailed Correlation was the particular test selected for each state read and other test analysis because it offered more stringent levels for rejecting the null hypothesis (H_0). Because the number of individuals whose scores were used in the study was 278, the confidence in the correlation was high.

Additionally, the study used the *t*-test to see if a statistically significant difference occurred between the two genders on the Nebraska Reading Comprehension/vocabulary Proficiency Exam. And, the study used a *t*-test to see if a statistically significant difference occurred between participants in the

Free and/or Reduced Lunch Program and the non-participants in the program on scores for the Nebraska Reading Comprehension/vocabulary Proficiency Examination.

The correlations between the locally designed CRTs and the Nebraska Reading Comprehension/vocabulary Proficiency Examination were of primary focus of the research. Much time and thought had gone into the creation of the local examinations, and the Language Arts teachers, specifically the American literature teachers, felt a need to see if a positive, reasonable correlation existed between the two reading tests, state and local, and if a positive, reasonable correlation existed between the Total CRT and the Nebraska Reading Comprehension/vocabulary Proficiency Exam. Were the locally designed tests measuring the knowledge deemed important by the Nebraska Department of Education as demonstrated through the Nebraska Reading Comprehension/vocabulary Proficiency Examination?

Correlation between Semester 1, Quarter A Total CRT (s1qatot), and Semester 1 Quarter A cold-read (s1qaread) and the state (state) reading results. Table 1 displayed that the correlation for the Semester 1 Quarter A Total CRT scores and the state read scores was a positive, two-tailed correlation of [r (123) = .546**, p< .01]. The positive, two-tailed correlation of the Semester 1 Quarter A cold-read scores and the Nebraska Reading Comprehensive/vocabulary Proficiency Examination scores was a positive correlation of [r (123) = .594**, p <.01.

The double asterisk (**) indicates the level of significance to be at or less than the .01 level.

TABLE 1

Correlation for Semester 1 Quarter A: Total CRT and Cold-read; and Total CRT and State Reading Examination.

Semester 1, Quarter A

Total CRT	Cold-read	State Reading
1		.546**
	1	.594**

** denotes a correlation of p<.01 level of confidence

Correlation between Semester 1, Quarter B Total CRT (s1qbtot), and Semester 1, Quarter B cold-read (s1qbread) and the state (state) reading results.

Table 2 displayed the positive correlations for the Semester 1 Quarter B Total CRT scores and Nebraska Reading Comprehension/vocabulary Proficiency Examination scores was a positive, two-tailed correlation of [*r* (123) = .566**, p< .01], and the correlation of the Semester 1 Quarter B cold-read scores and the state reading scores was a positive correlation of [*r* = .624**, *p,* .01].

The double asterisk (**) indicates the level of significance to be at or less than the .01 level.

TABLE 2

Correlation for Semester 1 Quarter B: Total CRT and Cold-read; and

Total CRT and State Reading Examination

Semester 1, Quarter B

Total CRT	Cold-read	State Reading
1		.566**
	1	.624**

** denotes a correlation of p<.01 level of confidence

Correlation between Semester 2, Quarter A Total CRT (s2qatot), and Semester 2, Quarter A, cold-read (s2qaread) and the state reading (state) results.

Table 3 displayed the correlations for Semester 2 Quarter A: a correlation for the Semester 2 Quarter A Total CRT scores and the state read scores was a positive, two-tailed correlation of [r (155) = .708**, p < .01], and correlation of the Semester 2 Quarter A cold-read scores and the Nebraska Reading Comprehension/vocabulary Proficiency Examination scores was a positive, two-tailed correlation of [r (155) = .627**, p < .01].

The double asterisk (**) indicates the level of significance to be at or less than the .01 level.

TABLE 3

Correlation for Semester 2 Quarter A: Total CRT and Cold-read; and

Total CRT and State Reading Examination

Semester 2, Quarter A

Total CRT	Cold-read	State Reading
1		.708**
	1	.627**

** denotes a correlation of p<.01 level of confidence

Correlation between Semester 2 Quarter B Total CRT (s2qbtot), and Semester 2, Quarter B cold-read (s2qbread) and the state (state) reading results.

Table IV displayed the correlations for Semester 2 Quarter B: a correlation for the Semester 2 Quarter B Total CRT scores and the Nebraska Reading Comprehension/vocabulary Proficiency Examinations scores was a positive, two-tailed correlation of [r (155)= .581**, p < .01], and the positive, correlation of the Semester 2 Quarter B cold-read scores and the Nebraska Reading Comprehension/vocabulary Reading Proficiency Examination scores was a positive, two-tailed correlation of [r (155)= .627**, p < .01].

The double asterisk (**) indicates the level of significance to be at or less than the .01 level.

TABLE 4

Correlation for Semester 2 Quarter B: Total CRT and Cold-read; and

Total CRT and State Reading Examination

Semester 2, Quarter B

Total CRT	Cold-read	State Reading
1		.581**
	1	.627**

** denotes a correlation of p<.01 level of confidence

The *t*-test comparison of mean scores by gender on the state reading produced significant results.

Table 5 displayed the *t*-test results for gender analysis on the Nebraska Reading Comprehension/vocabulary Proficiency Examination. The prediction was that no significant difference occurred between genders, and the *t*-test was a validation of the similarity of the means of the male and female results on the Nebraska Reading Comprehension/vocabulary Proficiency tests. The actual *t*-test rejected the null hypothesis H_0. [$t\ (251.66) = -2.23$, $p < .03$] level of confidence, and indicated that these mean scores did not occur through chance.

A significant difference in mean scores did occur for males and females, and this indicated that the eleventh grade females at this Midwestern, Class III, Class A school district were statistically more successful on the Nebraska Reading Comprehension/vocabulary Proficiency Examination than the eleventh grade males at this same Midwestern, Class III district, operating a Class A high school.

TABLE 5

Comparison of mean scores of Males and Females on the state reading examination

Gender	N	Mean	Std. Deviation
Male	138	34.9203***	8.80615
Female	140	36.9857***	6.48183

***denotes significance p.<03 level of confidence

The *t*-test for comparison of mean scores by participants and non-participants in the Free and/or Reduced Lunch Program on the state reading tests produced significant results.

Table 6 displayed the *t*-test results for socioeconomic grouping analysis. This grouping analysis was based on the students' participation or non-participation in the Free and/or Reduced Lunch Program. The prediction was that no significant difference occurred between participants in the free or reduced lunch program and non-participants in the program, otherwise known as full-paying lunch students. The actual *t*-test was to validate the similarity of performance for both Free and/or Reduced lunch participants and non-participants. The actual Sig. (2-tailed) *t*-test results rejected the null hypothesis H_{0}. [t (278) = - 4.496, p <.01] level of confidence, indicating these scores did not occur through chance. A statistically significant difference in mean scores for Free and/or Reduced Lunch Participants and the Non-participants. And analysis indicated that Non-participants in the Free and/or Reduced Lunch Program scored significantly higher on the Nebraska Reading Comprehension/vocabulary Proficiency Examination than the Participants.

TABLE 6

Comparison of mean scores of Free and/or Reduced Lunch Program Participants and Non-participants on the state reading test

Lunch	N	Mean	Std. Deviation
Participants	92	34.9203**	8.80615
Non-Participants	186	36.9857**	6.48183

**denotes p.<.01 level of confidence

Summary

Chapter IV explained the researcher's use of the SPSS software to calculate the results of this research study. The chapter detailed the data analysis of the four sets of correlations regarding the locally constructed CRT made of two components: CRTs over the taught material, and cold-reads tests intended to measure reading comprehension. The results showed positive correlations with these above mentioned scores and the Nebraska Reading Comprehension/vocabulary Proficiency Exam scores for each of four quarters.

The correlations were a positive .50 or higher for each quarter, so the Midwestern, Class III district, operating a Class A public high school personnel can feel that as a group and/or district they are headed in a desirable direction to meet the mission of educating their students at a level expected by the Nebraska Department of Education.

Chapter V: Conclusion

Purpose of the study

The purpose of the retrospective, correlational research study was to determine whether the locally designed, state standards-based, American literature criterion-referenced testing is a valid and reliable method of assessing student achievement on the Nebraska state reading standards in eleventh-grade language arts. The research also discovered that a statistical significance, although not a practical significance requiring change in instruction, existed in mean scores between genders on the Nebraska Comprehension/vocabulary Exam, and a statistical significance in mean scores existed between Participants and Non-participants in the Free and/or Reduced Lunch Program.

Summary of Findings

The Total CRT and state reading correlations were a positive .50 or higher for each quarter. Semester 1 Quarter A produced a coefficient of .594**; Semester 1 Quarter B produced a coefficient of .566**; Semester 2 Quarter A produced a coefficient of .708**; Semester 2 Quarter B produced a coefficient of .581**. A positive correlation of .50 demonstrated a good correlation. A positive correlation of .70 or higher for each of the correlations indicated a high correlation between the two variables.

The locally constructed reading tests and the Nebraska Reading Comprehension/vocabulary Proficiency Examination correlation was a positive .50 or higher for each quarter. Semester 1 Quarter A produced a correlation .546**; Semester 1 Quarter B produced a correlation of .624**; Semester 2

Quarter A produced a correlation of .708**; Semester 2 Quarter B produced a correlation of .627**.

The *t*-test of means of males and females for the Nebraska Reading Comprehension/vocabulary Proficiency Examination elucidated that the males did less well on the state reading test than the females, but the difference, although statistically significant, is not large enough to effect change in teaching, course structure, or scheduling. This conclusion was made at a p = <.03 level of confidence.

The *t*-test of means of Participant and Non-Participants in the Free and/or Reduced Lunch Program elucidated a statistically significant difference in performance on the Nebraska Reading Comprehension/vocabulary Proficiency Examination. The Participants in the Free and/or Reduced Lunch Program did less well than the Non-participants. This conclusion was made at a p = <.01 level of confidence. The difference in performance is large enough to warrant examination of the results with an eye to implementing assistance for these students.

Discussion

Although the correlations were all positive and significant at the *p* = <.01 level, initial observation notes that there was a difference between Semester 1 scores and Semester 2 scores. Why this difference occurs bears some discussion, speculation, or explanation.

Accounting for differences in semester results involves the following:

1. The fall semester, or first semester, of the 2009-2010 school year was the first time all the Language Arts teachers in the 9-12 program used the newly purchased grade level texts. This was true for the American literature teachers, as well. Teachers did not have the same familiarity of the material in the new texts as they had with the materials from previously used texts.
2. The first semester of the school year does not include any state tests for the eleventh grade students. So, it is imaginable that the focus for doing well was not present to the same degree in these students as those students in second semester Language Arts, specifically American Literature.
3. In block schedule, it is possible that students may take their required one semester of Language Arts in second semester of a year and then take the next year's required Language Arts or next level in first semester of that year, thus, making the Language Arts learning more closely connected. Conversely, students may take the one required semester of Language Arts in the first semester of one year, but not take the next year or next level of required Language Arts until the second semester of the following school year, thus making the consistency of learning opportunity uncertain.
4. The second semester of the school year 2009-2010 presented the American Literature teachers with all new students, but it was the second time through the text, support material, and CRT testing for the

teachers. This made the teachers more familiar with the material and presented an opportunity to be more thorough in their teaching.

5. The state of Nebraska eleventh grade writing exam occurred in Semester 2 Quarter A. This focus on state testing, the reinforcement of the importance of doing well, the knowledge that the students who do not receive a state writing score that meets the district standard would be scheduled for a writing course for remediation to meet graduation requirements may have had a positive effect upon the students' focus on learning and testing during the quarter.
6. The Semester 2 Quarter A calendar was extended by 20 minutes per day for the entire quarter, and no scheduled days away from school occurred. Perhaps the extra time at school or days at school, because of the calendar, caused the students to focus on better performance in testing.
7. The Nebraska Reading Comprehension/vocabulary Proficiency Exam occurred in Semester 2 Quarter B. Perhaps knowing the test was going to be administered affected preparation in both Quarters by the teachers and by the students.
8. Semester 2 could have had a larger population of students who are more Language Arts oriented, better test takers, or more mature as students.

Implications of the Study

The implications of this study appeared when the relationship between the district and state scores was revealed; the relationship between the scores will allow the local school district the opportunity of any of the following:

1. The district does not need to reconstruct the curriculum and the subsequent summative testing in its entirety because of low correlations or negative correlations between the curriculum knowledge scores and the reading scores; or
2. The district does not need to make wide scale adjustments to the measuring tools or curriculum because analysis shows neither of these to be negative factors through low or negative correlations; or
3. The district does not need to make a large-scale adjustment to the teaching staff, methodology, or selected readings used to teach eleventh grade students because no low or negative correlations between locally constructed tests and the state reading test were discovered; or
4. The district does not need to adjust the rigor of the reading opportunities presented for student learning because low or negative correlations were discovered; or
5. The district can set its sights each succeeding year toward reaching a high, positive correlation between its locally constructed tests and the state reading exam and continue its on-target work of successfully educating students.

Recommendations for the Present

Standards-based testing was intended to determine the relationship of student scores between those of the locally constructed, standards-based, criterion-referenced test (CRT) for 278 eleventh grade students at a Midwestern, Class III district, operating a Class A high school, and the scores of those same students on the Nebraska Reading Comprehension/Vocabulary Proficiency Examination for eleventh grade, public high school students. This study determined those district created tests positively correlated between the exams in their entirety with the Nebraska Reading Comprehension/vocabulary Proficiency Examination, and there was a positive correlation between the cold-read portion of the CRT and the Nebraska Reading Comprehension/vocabulary Proficiency Exam. Correlations showed the relationship between student scores for the entire number of eleventh grade students who had taken all three tests, and for sub-sets of students: first semester students Quarters A and B, second semester students Quarter A and B, male students and female students, Free and/or Reduced lunch Participating and Non-participating students.

In the rush to give students the opportunity to work in the format of cold-read tests, and for these same students to experience being tested for reading comprehension and vocabulary proficiency, and for these same students to be prepared for reading test that was to be administered by the state, the local Midwestern, Class III district, operating a Class A high school did not statistically evaluate the cold-reads that were used on the locally devised CRT Quarter examinations. The district should run statistical analysis on these tests. The

cold-read format was modeled to match that of the state test, but other concerns were not considered or recognized. It would seem imperative to have a readability score completed on the cold-read selections used for quarterly evaluation of eleventh grade students.

As with any objective test during its first year of use, inadvertent errors are bound to occur. The locally designed Quarter CRT examinations should be thoroughly scrutinized by using item analysis for poor choices and detractors, by searching for poor test questions, and evaluating page design to detect any distracting presentation on the page.

A final recommendation for the next administration of these examinations is a better method of tracking students who have not taken a quarter examination or the state examination as some of the students in 2010. Without this student information, analysis of test results is incomplete, and district direction and a general plan for assisting these students is not as thorough as it might be, and information needed to improve the testing for the next group of eleventh grade students is missing.

Recommendation for Future Research

Future research on eleventh grade students over the next seven years, the usual lifetime of a text series, is needed and would help to clarify the value and effect of teacher familiarity with the material and the effect it has on testing.

Future research on vocabulary usage alone would be informative. The eleventh grade students for the school year 2010-2011 will have been formally taught Latin and Greek roots in their Language Arts, specifically English 10

Literature, classes. Determining the effect or the effectiveness of such teaching could be valuable.

Future research that finds the relationship between locally designed, cold-read tests, the state reading tests, and reading scores for college entrance examinations could be beneficial for students and the school district that produces college bound students. This research could be retrospective and include the statistical results for this last school year 2009-2010 as a taking-off point.

Future research regarding the relationship of cold-read test scores in Language Arts grades 9 and 10 and the scores on Quarter CRT Total tests for those grade levels should be done to assess the reading level used in the test questions and cold-read selections. When this analysis is complete, credibility is established regarding the test results, and this creditability allows educators to design and conduct interventions for needy students earlier in the student's high school experience.

Future research regarding the relationship of reading quiz scores, for quizzes occurring before discussion, and summative tests scores over the same reading assignments could be valuable as another method of cold-read preparation for comprehension and vocabulary proficiency.

Summary

Research regarding reliability and face validity of high school standards-based tests, including subject area and general reading, must continue so information can be ascertained whether or not these tests are reliable measures

of subject knowledge, reading comprehension, and vocabulary proficiency. The Nebraska Department of Education, Nebraska school districts, teachers, parents, and students must be conscientious regarding their awareness of state standards and about making a consistent effort in meeting or exceeding those standards to ensure Nebraska students are armed with the proper tools to help them succeed in the educated world.

References

Bond, L. A. (1996). Norm-and criterion-referenced testing. *Practical Assessment, Research & Evaluation, 5*(2). Retrieved March 3, 2009 from http://PAREonline.net/getvn.asp.

Bordignon, C., & Lam, T., (2004). The early assessment conundrum: Lessons from the past, implications for the future. *Psychology in the Schools, 41*(7), 737-749.

Bryant, M.T. (2004). *The portable dissertation*. Thousand Oaks, CA: Corwin Press.

Budnik, T., (2001). An estimate of the reliability of a technique of increasing educational accountability through goal analysis involving community, staff, and students. *Journal of Educational Research, 71*(5) 251-261.

Christ, T., & Schanding, G. (2007). Curriculum-based measures of computational skills: A comparison of group performance in novel, reward, and neutral conditions. *The School Psychology Review, 36*(1), 147-58.

Crowley, J., & Hauser, A., (2007). Evaluating whole school improvement models: Creating meaningful and reasonable standards of review. *Journal of Education for Students Placed At Risk, 12*(1), 37-58.

Dappen, L., & Isernhagen, J., (2005). *Nebraska STARS: Assessment for learning. planning and changing. 36*(3&4),147-156.

DiRanna, K., Osmundson, E., Topps, J., & Gearhart, M., (2008). Reflections on assessment. *Principal Leadership, 9*(1), 22-27.

Downing, S., & Haladyna, T., (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*(1), 61-82.

Espin, C., Scierka, B., Skare, S., & Halverson, N., (1999). Criterion-related validity of curriculum-based measure in writing for secondary school students. *Reading & Writing Quarterly* (serial online), *15*(1), 5-27.

Espin, C., Shin, J., & Busch, T. (2005). Curriculum-based measurement in the content areas: Vocabulary matching as an indicator of progress in social studies learning. *Journal of Learning Disabilities, 38*(4), 353-63.

Espin, C., Wallace, T., Campbell, H., Lembke, E., Long, J., & Ticha, R. (2008). Curriculum-based Measurement in writing: predicting the success of high-school students on state standards tests. *Exceptional Children, 74*(2), 174-93.

Fast, E.F. (2004). Evaluation report on the Nebraska state department of education's district assessment portfolio training process. (2004, October 11). Nebraska Department of Education. Retrieved March 12, 2010, from http://www.nde.state.ne.us/assessment/techreports/.../EFFreportonDAP_10-12-04. pdf.

Ferrandino, V., (2006). Getting a better perspective on testing. *Principal, 85*(4), 4.

Fien, H., Baker, S., Smolkowski, K., Smith, J., Mercier, J., Kame'enui, E., Beck, E., & Thomas, C. (2008). Using nonsense word fluency to predict reading proficiency in kindergarten through second grade for English learners and native English speakers. *The School Psychology Review, 37*(3), 391-408.

Foote, M. (2007). Keeping accountability systems accountable. *Phi Delta Kappan, 88*(5), 359-63.

Frisbie, D., Miranda, D., & Baker, K. (1993). An evaluation of elementary textbook tests as classroom assessment tools. *Applied Measurement in Education, 6(*1), 21-36.

Gallagher, C. (2004). Turning the accountability tables: Ten progressive lessons from one 'backward' state. *Phi Delta Kappan, 85*(5). 352-360 Retrieved March 29, 2009, from http://www.pdkintl.org/kappan/k0401gal.htm.

Haladyna, T., Haas, N., & Allison, J. (1998). Continuing tensions in standardized testing. *Childhood Education, 74*(5), 262-73.

Herman, J., Gearhart, M., & Baker, E. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment, 1*(3). 201-224.

Hintze, J., & Christ, T., (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *The School Psychology Review. 33*(2), 204-17.

Huitt, W., (1999). Implementing effective school achievement reform: Four principles. Paper presented at the School Counseling Summit, Valdosta State University, Valdosta, GA. Retrieved July 28, 2009, from http://chiron.valdosta.edu/whuitt/files/school_reform.html.

McGehee, J. & Griffith, L., (2001). Large-scale assessments combined with curriculum alignment: Agents of change. *Theory into Practice, 40*(2). 137-144.

Nebraska School Activities Association - Football Classification. Retrieved March 10, 2010 from http://nsaahome.org/textfile/fbl/fbclass.

Ogawa, R., (2003). The substantive and symbolic consequences of a district's standards-based curriculum. *American Educational Research Journal, 40*(1), 147-176.

Pennington, M., (2009, January 10). How to differentiate reading fluency practice. *Pennington Publishing Blog.* Retrieved June 29, 2010 from http://penningtonpublishing.com/blog/reading/how0to-differentiate-reading-fluency-practice.

Rivera, J., (2005). Designing assessments – where does alignment fit in? *The Agricultural Education Magazine. 78*(1), 15-16.

Roschewski, P., Isernhagen, J., & Dappen, L., (2006). Nebraska STARS: achieving results. *Phi Delta Kappan. 87*(6), 433-437.

Schultz, R., (2000). Teachers as learners: Studying a three-phased rubric assessment plan, *Gifted Child Today, 25(*4), 38-45, 65.

Sloane, F. C., & Kelly, A. E., (2003). Issues in high-stakes testing programs, *Theory Into Practice, 42*(1), 12-17.

Thattai, D., (2001). A history of public education in the united states, A*idmn-Ejournal, 1-5.* Retrieved July 9, 2010 from http://www.servintfree.net/~aidmn-ejournal/publictions/2001-11/PublicEducationaInThe United States.html.

Vu, P., (2007). States gaming NCLB system, *The PEW Center on the States,* Nov. 13, 2007, 1-4. Retrieved March 3, 2010 from http://www.stateline.org/live/dteails/story?contentid=256971

Zhang, Y., (2009) State high school exit exams: Trends in test programs, alternate pathways, and pass rates, *Center on Education Policy,* Nov. 5, 2009, 2. Retrieved March 3, 2010 from http://www.cep-dc.org/index.cfm?fuseaction=page.viewPage&pageID=579&nodeID=1

Appendix A: Data Collection

Spreadsheet Categories and Abbreviations

Semester 1 Quarter A CRT (s1qalit),

Semester 1 Quarter A cold-read (s1qaread),

Semester 1 Quarter A Total (s1qatot),

Semester 1 Quarter B CRT (s1qblit),

Semester 1 Quarter B cold-read (s1qbread),

Semester 1 Quarter B Total (s1qbtot).

Semester 2 Quarter A CRT (s2qalit),

Semester 2 Quarter A cold-read (s2qaread),

Semester 2 Quarter A Total (s2qatot),

Semester 2 Quarter B CRT (s2qblit),

Semester 2 Quarter B cold-read (s2qbread),

Semester 2 Total (s2qbtot),

State Reading (stateread),

Gender (sex),

Free and/or Reduced Lunch Participants and Non-participants (lunch).

Appendix B: Consent to do Research

Xxxxxxx Xxxxxx Xxxxxxx

XXXX XXXXXX XXXXXXXXX X XXXXXXXXXXXXXX XXXXXX

XXX XXXX Xxx XXXXXX

XXXXXXX, NEBRASKA

PHONE XXXxXXXxXXXX

January 20, 2010

Graduate School

College of St. Mary

Omaha, Nebraska

To Whom It May Concern:

This letter was written as authorization for Ms. Perk Beckman to conduct research utilizing Criterion Referenced Test data from the eleventh grade student population at Xxxxxxx Xxxx School, for comparison with and to Nebraska State Reading Test results. Information of a personally identifiable nature concerning individual students as it relates to federally protected income and free-reduced lunch information cannot be provided. However, performance data for those individuals can be provided in the aggregate, as determined and approved by XXx Xxxxx Xxxxxx, Executive Director of Curriculum, Instruction and Assessment.

It is preferred that reference to Xxxxxxx Public Schools be made within any reports as, "a Class Three School District, operating a Class A High School," or some variation of that descriptor.

Should further information be required, please contact me at your convenience.

Sincerely,

Xxxxxxx Xxxxxx, Superintendent

Xxxxxxx Public Schools

Appendix C: IRB Approval

April 24, 2010

Ms. Priscilla Beckmann
College of Saint Mary
7000 Mercy Road
Omaha, NE 68106

Dear Ms. Beckmann,

The College of Saint Mary Institutional Review Board has reviewed your revised research proposal **State of Nebraska Reading Comprehension/Vocabulary Test and the Local American Literature Criterion-referenced Test for Public School Eleventh Grade Students: A Correlational Study** at the exempt level. Approval has been granted based on appropriate changes and corrections being made in your proposal.

Your official research number is #CSM 10-10. This should be used in all correspondence regarding your project. At the conclusion of your project, please submit the "Closing the Study" form, which appears on page 40 of the revised IRB Manual (available on the IRB Community site which appears on the main CSM website, after clicking "mycsm").

Please let me know if I may be of further assistance. Good luck with your research!

Sincerely,

Sue Schlichtemeier-Nutzman

Dr. Sue Schlichtemeier-Nutzman
Assistant Professor
(402) 416-8599 Office Cell

E 68106-2606 • 402.399.2400 • FAX 402.399.2341 • www.csm.edu