AN EDUCATOR'S PERCEPTION OF STARS FROM SELECTED NEBRASKA PRINCIPALS

By

Philip B. Warrick

Presented to the Faculty of

The Graduate College

at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Education

Major: Educational Administration

Under the Supervision of Professor Larry L. Dlugosh

Lincoln, Nebraska

December 2005

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

DISSERTATION TITLE

AN EDUCATOR'S PERCEPTIONOF STARS FROM SELECTED NEBRASKA PRINCIPALS

BY

PHILIP B. WARRICK

SUPERVISORY COMMITTEE:

Approved — Date

Larry L. Dlugosh — 10/17/05
Signature

Larry L. Dlugosh
Typed Name

Jody C. Isernhagen — 10/17/05
Signature

Jody C. Isernhagen
Typed Name

Donald F. Uerling — Oct 17, 2005
Signature

Donald F. Uerling
Typed Name

L. James Walter — 10/17/05
Signature

L. James Walter
Typed Name

Martha Bruckner — 10-17-2005
Signature

Martha Bruckner
Typed Name

Signature

Typed Name

UNIVERSITY OF Nebraska Lincoln

AN EDUCATOR'S PERCEPTION OF STARS FROM SELECTED NEBRASKA PRINCIPALS

Philip B. Warrick, Ed.D.

University of Nebraska-Lincoln, 2005

Adviser: Larry Dlugosh

This study was part of a larger four-part study that investigated the problem of implementing a state accountability system through locally developed assessment and reporting of student performance. The purpose of this quantitative, descriptive study was to survey and describe the perceptions of Nebraska principals regarding the implementation of the School-based, Teacher-led Assessment Reporting System (STARS).

Seven hundred eighteen educators from across the state of Nebraska participated in this four-part research study. Upon conclusion of the four studies, the researchers completed a comparative analysis of the four groups. The major findings of the four-part study was that educators were generally positive in their perceptions of STARS. ESU staff developers gave the STARS model the most positive responses of the four educator groups surveyed; conversely, principals gave STARS consistently lower marks relative to the other groups. Scores from assessment coordinators and principals were generally similar and usually placed between the scores of ESU staff developers and teachers.

All groups perceived that public education in Nebraska improved due to STARS with results aligning to educators' general perceptions of STARS. Other major findings of the four-part study were that curricular and assessment knowledge improved due to STARS as did teachers as leaders of learning in this teacher-led assessment system.

ACKNOWLEDGEMENTS

I dedicate this dissertation and my doctoral degree to my wife Tory and my daughters Lauren and Madison. I have succeeded because the three of you were unselfish and willing to give me time and support to pursue my degree. It is my intention to support all of you as you pursue your dreams and goals in life. While this degree allows me to be called doctor, the three of you allow me to be called dad.

During my educational journey I have made a professional and personal friend in my advisor Dr. Larry Dlugosh. His positive attitude and ability to constantly encourage people has always inspired me. It is my hope to do the same for others in the future as he has done for me.

TABLE OF CONTENTS

List of Tables

List of Appendices

CHAPTER 1

INTRODUCTION

Schools of the 21st century operated in an environment of accountability brought forward by public discontent, political interests, and concerns about international competition (Chatterji, 2002; Marzano & Kendall, 1998). These concerns accounted for many national educational movements. The most recent round of national accountability was noted by the reauthorization of The Elementary and Secondary Education Act in 2002. This legislation required each state to submit to the United States Department of Education an accountability plan called the Consolidated State Application. As national accountability set the stage, local states were left to outline plans spelled out in the principles required by No Child Left Behind (See Appendix A) in 2002 (No Child Left Behind Act, 2002).

Forty-eight states in the country responded to the accountability age by developing "statewide mandated tests" (Bandalos, 2004, p. 6). Nebraska adopted an approach different than any other state regarding educational accountability. Dr. Doug Christensen, the Commissioner of Education, along with the state legislature, reached agreement in 1999 to use an assessment system not based upon a single state standardized test but on multiple measures designed by each of the state's local school districts (Boss, Endorf & Buckendahl, in press; Roschewski, 2004). Nebraska's system represented what Jones and Ongtooguk (2002) described as "other possibilities for assessing student learning and holding schools accountable that go beyond the use of a single high-stakes test" (p. 503). The Nebraska School-based, Teacher-led Assessment and Reporting

System (STARS) was unique from systems in other states in that it was internally focused with its foundation at the local district level.

To add to the research on this topic, the focus of this quantitative, descriptive study was to determine and describe the perceptions and practices of educators about the implementation of the Nebraska STARS. As part of this system, schools were required to administer locally developed assessments to measure the academic content standards and report the student achievement results to the Nebraska Department of Education (NDE). Since this model was a bottom-up approach, the perceptions of educators about STARS was paramount to an analysis of the system (NDE, 2002).

In partnership with the Buros Institute for Assessment Consultation and Outreach (BIACO), hereafter referred to as Buros, the state department of education developed a set of criteria to determine the quality of the local assessment system. The schools were required to submit a District Assessment Portfolio (DAP) that outlined the procedures used to meet the Six Quality Assessment Criteria (NDE, 2002). The district portfolios were evaluated by outside assessment experts to determine the technical quality of the local assessment systems. The student performance and assessment quality results were published as part of the State of the Schools Report. Moreover, state accreditation regulations, NDE Rule 10, mandated that schools comply with the provisions of the state assessment system (NDE, 2004).

STARS was a relatively new and unique assessment approach during an era of accountability in education. From the commissioner of education to the classroom teacher, the experience of developing and implementing STARS changed the way many educators conducted business. To this end, it was important to share the perceptions and practices of educators about the implementation of the Nebraska STARS.

Statement of the Problem

This study was part of a larger four-part study by Toby Boss, Daniel Endorf, Tamara Heflebower, and Philip Warrick. The research team investigated the problem of implementing a state accountability system through locally developed assessment and reporting of student performance.

Purpose Statement

The purpose of this quantitative, descriptive study was to survey and describe the perceptions of Nebraska principals regarding the implementation of STARS. The other parts of the larger study included a study of educational service unit (ESU) staff developers perceptions by Tamara Heflebower, a study of teacher perceptions by Dan Endorf, and a study of assessment coordinators perceptions by Toby Boss.

Research Questions

Three research questions guided the four-part study: The term "educators" encompassed the four different groups that were studied. In this particular study the term "educators" represented principals. The three research questions included:

1. What are the perceptions of educators about STARS as it related to education in Nebraska?
2. What are the perceptions of educators about the curriculum, instructional, and assessment practices used to implement STARS in Nebraska?

3. What are the perceptions of educators about the impact of STARS on the professional abilities of educators across the state of Nebraska?

Theoretical Framework

Design of the survey instrument was tailored to fit the perceptions and practices of schools within the context of curriculum, instruction and assessment theory. The theories were derived from current research in the field.

Standards and Assessment Theory

The basis for the curriculum theory evolved from Marzano (2003), who researched factors influencing student achievement. From this meta-analysis of research about effective schools, Marzano concluded that a guaranteed, viable curriculum was the top school-level factor related to student achievement. In order to implement a guaranteed and viable curriculum, Marzano proclaimed that schools should identify the essential curriculum, guarantee it can be taught in the time available, and organize the curriculum to allow students the opportunity to learn it. Once the curriculum was in place, Marzano stressed the need to guarantee that teachers implemented the essential curriculum, and that schools protected instructional time.

Instructional theory related to curriculum theory. Mid-continent Research for Education and Learning (McREL) analyzed and recorded instructional strategies available to teachers (Marzano, Pickering, & Pollack, 2001). Using meta-analysis research from the past 30 years, Marzano concluded that nine instructional strategies were critical to teacher effectiveness. The nine instructional strategies were: (a) identifying similarities and differences; (b) summarizing and note taking; (c) reinforcing effort and providing recognition; (d) homework and practice; (e) nonlinguistic representations; (f) cooperative learning; (g) setting objectives and providing feedback; (h) generating and testing hypotheses; and (i) questions, cues and advanced organizers.

Schools in Nebraska were allowed to design and use locally based assessments, but they were also required to evaluate the technical quality of the assessments based upon a set of six quality criteria. Buros developed the six criteria in collaboration with NDE. The six quality criteria were: (a) assessments match the standards; (b) students have an opportunity to learn the content; (c) assessments are unbiased and sensitive to cultural differences; (d) assessments are at an appropriate level; (e) assessment scores or decisions are reliable; and (f) mastery levels are set appropriately (Plake, Impara, & Buckendahl, 2004).

Assumptions

The first assumption of the study was that school districts were attempting to meet the requirements as set forth by Nebraska Department of Education Rule 10. Rule 10 included all of the NDE school accreditation requirements, such as course offerings, teacher certification, and instructional units per year. Under Rule 10, schools were required to comply with the provisions of the state assessment system. Rule 10 required that schools not only meet the procedural requirements of the assessment system, but also improve low ratings on student performance and/or assessment quality (NDE, 2004). Accreditation was assumed to be important to school districts in the state, and school leaders should have undertaken measures to meet these regulations.

A second assumption was that STARS affected the curriculum alignment, instructional delivery, and assessment practices of classroom teachers in the state. The six quality assessment criteria outlined the components that districts included in the documentation of the local assessment system through the District Assessment Portfolio (DAP) (Plake et al., 2004). This portfolio requirement caused school districts to examine and refine classroom practices, forcing educators to do business in a different manner than before STARS.

Delimitations

A delimitation of the study was that the sample included educators who had experience implementing STARS. At the time of this study, the STARS regulations only included the subjects of reading and mathematics, and only grades 4, 8, and 11 in many cases. This factor delimited the study to educators who have directly experienced the STARS regulations within their own district.

Limitations

As the state's regulatory agency, NDE worked with schools on a variety of topics, from assessment to funding to reorganization. Since STARS functioned under the NDE umbrella, the potential existed for respondents to have a skewed perception of NDE. Therefore, one limitation of the study was the overall perception of NDE by sample respondents.

A second limitation of the study was individual district philosophies about the STARS system. Gallagher (2004) stated that the success of STARS within a school system depended upon "For whom the work is being completed" (personal communication, August 25, 2004). Those districts holding the philosophy that STARS furthered educational purposes were more positive toward STARS than their counterparts who viewed it as another state mandate. Therefore, the philosophy a district held toward STARS may have impacted the answers of individual respondents from that district.

A third limitation of the study was the reliance on the respondents to be truthful in their responses. The anonymity of the respondents was protected, yet a common problem in this type of study was the desire of the respondents to provide the socially desired answers (Dillman, 2002).

A fourth limitation of the study was the technological competency of the sample. Educators with limited technology skills may have chosen not to, or been unable, to participate in the survey.

A fifth limitation of the study was the availability of accurate and reliable e-mail addresses for educators in the state of Nebraska. There was no single, statewide database for obtaining these e-mail addresses. In fact, due to the advent of spam, organizations were strictly protecting e-mail addresses making them more difficult to obtain.

Definition of Terms

Accountability—The concept of educators being held responsible for student achievement that meets a given set of standards. Achievement is demonstrated by showing mastery on a norm-referenced or criterion-referenced test (National Forum on Assessment, 1995).

Assessment—The process of gathering information about student achievement at both the large-scale standardized and classroom levels to make instructionally relevant decisions (Stiggins, 1995).

Buros Institute for Assessment Consultation and Outreach (BIACO)—The independent evaluator of the assessment practices employed by the STARS system.

Criterion-referenced Assessments—Assessments in which each student's score is compared to a preset level of acceptable performance rather than being compared to a norm group for interpretation (NDE, 2001; Stiggins, 1995).

District Assessment Portfolio—A summary of the assessment practices each school district utilized in the implementation of STARS. Districts are required to submit portfolios and receive a rating from independent reviewers (NDE, 2004).

Educational Service Units (ESU)—Intermediate support agencies that provide support to schools on a regional basis; there are 17 regional ESUs in Nebraska.

Nebraska Department of Education (NDE)—Nebraska's regulatory agency for public education. The concept of STARS originated within this agency.

No Child Left Behind (NCLB)—A piece of legislation enacted by the 107th Congress to close the achievement gap with accountability, flexibility, and choice so that no child is left behind (United States Department of Education, 2002).

Norm-referenced Tests—are defined as assessments that are used to sort or rank students along a continuum of achievement. The students are compared to a norm group of students who took the test under the same conditions for interpretation (NDE, 2001; Stiggins, 1995).

Six Quality Assessment Criteria—The six components to quality assessment practices as specified by the BUROS Institute of Educational Measurement. The six quality criteria include the following:

1. assessments match the standards,
2. students have an opportunity to learn the content,
3 assessments are unbiased and sensitive to cultural differences,
4. assessments are at an appropriate level,
5. assessment scores or decisions are reliable, and
6. mastery levels are set appropriately. (Plake et al., 2004)

School-based Teacher-Led Assessment and Reporting System (STARS)—Nebraska's procedure used to match local standards to state standards: writing assessments over those standards, giving the assessments, recording and reporting the results, and preparing a portfolio explaining the whole process for the state (NDE, 2001).

Significance of the Study

Several studies added to the body of research regarding the STARS system. Gallagher (2004) and Isernhagen and Dappen (2004) studied the process of collection and usage of STARS data. Other studies focused on the development and implementation of assessments—first, Roschewski (2002) from the district perspective, and then Heusman (2004) from the lens of the classroom teacher. Specht (2004) and Weichel (2002) connected the attitudes of educators about STARS through a study of the attitudes of principals, while Beran (2003) concentrated on the attitudes of classroom teachers.

Furthermore, Bandalos (2004) conducted a focus group study about the attitudes of both teachers and principals.

This study added to the body of research in multiple ways. First, the study had significance because it examined the perceptions of a broad base of public educators (teachers, principals, assessment coordinators, and ESU staff developers) regarding curriculum, instruction, assessment, and the STARS process in a distinctively different manner. The sheer diversity of the educational positions selected for the survey made this study significant to the growing body of research regarding STARS and assessment systems in general.

A second reason this study was significant related to the inherent differences associated with the Nebraska accountability and assessment model and other state-wide models across the country. A comprehensive study of the various educational positions across the state made the survey pertinent to decision-makers intrigued by alternative possibilities to statewide testing programs.

Another reason the study was significant related to concerns regarding the demands of NCLB. Concerns about the federal requirements regarding usage of statewide assessments were not only found in the educational arena, but also in the measurement community. Furthermore, colleges and universities should have had interest in this research as they considered curriculum revisions in the areas of assessment literacy, preparation, and data analysis for pre-service teachers and teachers enrolled in graduate school.

Policy makers took interest in STARS as "it was designed to avoid the pitfalls caused by misuse of information from standardized tests" (Roschewski, 2004, p. 9). If state policy makers were going to make decisions based on student achievement, the results should reflect what students learned rather than their inherent abilities and experiences so often indicative of norm-referenced assessment systems.

Finally, STARS was a relatively new system. NDE may have used this information to assist in the evaluation of their assessment model.

CHAPTER 2

REVIEW OF THE LITERATURE

The review of the literature utilized a blend of current journals, reports, and other pertinent articles to address the topic of study. The following narrative highlighted this in-depth review and is divided into six sections. The first section related to change and the role of change forces in education. The second section was devoted to the recent accountability movement at both the federal and state level. Following on the coattails of accountability was assessment; this section outlined current literature regarding the role of assessment in the field of education. The funnel narrowed even further with a discussion of STARS, the Nebraska accountability and assessment model. Following STARS was an in-depth analysis of how teachers, and later educational leaders, have handled the many changes associated with the accountability and assessment era.

Change

An Environment of Change

The word "change" was open to interpretation. An understanding of "change" depended largely on the situation and circumstance surrounding the change. Expert researcher Michael Fullan (2001) trumpeted the unknown characteristics of change when he stated,

> Change is a double-edged sword. The change process occurs at an inconsistent pace and raises the level of anxiety in everyone impacted by the change. Yet when things are unsettled, it is a prime opportunity to move forward and create breakthroughs not possible in stagnant societies. (p. 1)

The concept of change in the early 21st century evolved with technological advances and movement toward a global community. In turn, people dealt with change far more often than in previous generations (Slater, 2003). Kotter (1996) stated, "People of my generation or older, did not grow up in an era when transformation was common. The norm back then was stability" (p. 18).

Change operated under such names as restructuring, reform, and innovation. Regardless of nomenclature, change functioned in the form of a process. The change process was about people, emotion, and level of involvement. When a change occured in an organization, people compensated in different ways and at different paces. Some embraced change immediately and "jumped on board," while others resisted the change because of fear, concern, and/or the fact that their particular position would be different (Fullen, 2001; Kotter, 1996; Rogers, 1995; Slater, 2003).

By understanding how people reacted to change and the multiple stages of the change process, there existed the opportunity to be successful in a world of constant change, including the world of education.

Change and Education

Change and education can be traced back to the earliest days of formal education with Horace Mann. Debates about similar curricula, uniform textbooks, and teacher training all played a role in creating a system that ensured all students received a similar educational experience (Glickman, 2001; Lederman & Niess, 2000).

Various local, state, and federal forces have traditionally placed pressure on education to engage in some type of change movement. From the United States Supreme Court case, *Brown v. Board of Education* in 1954, which ended the legality of racially segregated schools, to the 1970's, when mandates for students with special needs were placed on schools with the passage of Public Law 94-142, public schools have oftentimes been reactionary to political winds of change (Gray, 1997; Meyerson, 2001).

The Modern Assessment and Accountability Movement

The winds of change regarding the modern assessment and accountability age in America began gusting in the 1980s. Political forces turned their attention toward the process of schooling and the academic performance of students in a 1983 report entitled, *A Nation at Risk* (National Commission on Excellence in Education, 1983). In regard to this report George and McEwin (1999) stated, "This shocking report charged public education with a level of inadequacy that rocked U.S. confidence in public education and shocked educators into a variety of responses that lasted more than a decade" (p. 12). Numerous defenders of public education emerged since the release of the report and offered data to show the report was both inaccurate and laden with political ideologies. Regardless of motive, the report remained one of the most influential change forces in the history of American education (English & Hill, 1990; George & McEwin, 1999; Pulliam & Van Patten, 1995).

In the years that followed *A Nation At Risk*, more than 30 additional investigations, reports, and recommendations came forth from educational change agents (Pulliam & Van Patten, 1995). One such document, the SCANS report, identified the public school as the primary source in the development of skills needed to obtain and maintain jobs (Secretary of Labor's Commission on Achieving Necessary Skills: SCANS Report, 1991).

Soon after the SCANS report, the National Association of Secondary School Principals (1996) collaborated with the Carnegie Foundation to release *Breaking Ranks: Changing an American Institution*. *Breaking Ranks* made many recommendations. One recommendation in particular spoke volumes about the changing landscape of assessment and accountability in public education:

> Accountability demands that a high school have a set of objectives and assess and report the extent to which they are met. The purpose of this accountability is to ensure that teaching and learning serve the needs of students to the fullest extent. (p. 53)

These words proved prophetic.

Other change agents also set their sights on the American educational system. Federal and state policymakers initiated policies aimed at improving education. Systemic reform of many varieties occurred across the nation that set academic standards for students, aligned policy with standards, and restructured educational governance to allow for shared decision-making at the local level (Glickman, 2001; Loup & Blasé, 1999). With the enactment of the NCLB of 2002, performance-based educational accountability became a federal mandate, trumping the decade of state activity and array of accountability systems (Hess, 2003; No Child Left Behind Act, 2002).

An environment of change saturated the public education arena in the early 21st century. While assessment and accountability were buzzwords in current educational lingo, few truly understood the complexity of the terms or the difference between them. While these terms intended to work together as concepts, assessment and accountability were different entities, each having an individual identity and purpose.

Accountability

History of the Accountability Movement

The history of the modern accountability movement can be traced back to 1983 with publication of *A Nation at Risk* (Marzano & Kendall, 1998). The National Commission on Excellence in Education (1983) recommended that schools adopt rigorous standards and that achievement tests occur as students move along the educational path. The panel also recommended that testing be part of a national program and administered by individual states.

Linn (2000) explained that in the 1970s and 1980s states began to adopt minimum competency testing (MCT), utilizing standardized tests to measure student achievement. The number of states using some kind of MCT rose dramatically during this timeframe. These tests created an inflated impression of student achievement, known as the Lake Wobegon Effect, in which almost all districts across the country reported student scores above the national average.

President George H. Bush called an education summit of the nation's governors in 1987 (Chatterji, 2002; Marzano & Kendall, 1998). Other meetings in subsequent years led to the creation of The National Goals Panel (1991), which was a federal publication with two specific goals connected with the modern standards movement. The two goals were to demonstrate competency in core subject areas, and to hold the highest scores in the world for mathematics and science. The work of this panel led to congressional action in 1994, with the adoption of Education Goals 2000 (Chatterji, 2002).

President Clinton met with the nation's governors for another summit on education in 1997. The primary accomplishment of this meeting was the creation of state standards for accountability, from which each state received the challenge of making the standards movement a success (Chatterji, 2002; Marzano & Kendall, 1998). The accountability movement entered into the civil rights arena the next year with publication of *A Nation Still at Risk* (The Center for Educational Reform, 1998), which contended that every child deserved an equal educational opportunity.

No Child Left Behind

The passage of the NCLB Act of 2002 took the accountability system to the federal level with the objective of reaching 100% student proficiency in math and reading by the spring of 2014 (NCLB, 2002). This plan required each state to submit a Consolidated State Application and receive approval from the United States Department of Education (USDE). The provisions of NCLB are outlined in the principles of the *Consolidated State Application Accountability Workbook* found in the Appendix A.

NCLB (2002) required that each state provide public reporting of a state report card with numerous required data elements, such as professional qualifications of all teachers, student achievement scores, high school graduation rates, and a comparison between the achievement levels of each student and the state's annual measurable objectives.

Linn, Baker, and Betebenner (2002) analyzed the implications of the NCLB approach to assessment and accountability. They wrote that Adequate Yearly Progress (AYP) was measured at the school level, and set expectations for rapid improvement for not only the total group, but also certain subpopulations of students. If schools did not make the state AYP objective two years in a row for either the total population or a subgroup, the school was identified for needs improvement. Indeed, standards-based accountability became the definition for educational reform, with the school as the basic unit of analysis (Chatterji, 2002).

All 50 states had an accountability plan approved under the NCLB guidelines. Individual state plans were summarized in Appendix B complete with a summary of the tests, subjects, and administration of the state accountability systems.

A review of each state-by-state accountability plan indicated that all but one state used state tests to determine proficiency on standards. However, the concept of the statewide test can be misleading. Many states used different tests to measure proficiency in certain content areas. Tests tended to be administered within specific timeframes, with several states allowing students multiple opportunities to demonstrate proficiency. Eleven states required passage of a state test as a requirement for graduation, and provided opportunities for students to take the test as early as tenth grade. Almost all states built accountability systems on state-developed tests, written by teachers from their state and based on the state's content standards, as opposed to nationally-normed standardized tests. Only The District of Columbia (SAT-9) and Iowa (ITBS and ITED) used norm-referenced assessments. All states measured language arts and math, while 36 states measured science and 23 states measured social studies.

New Accountability Systems

The passage of NCLB ushered in a new version of accountability. Fuhrman (2003) suggested that an accountability model like NCLB improperly assumed that performance can be accurately and authentically measured by the assessment instruments in use, that high-stakes testing motivated educators and students, and that higher levels of performance would result from the plan.

Policy makers typically had an affinity for assessment. Politicians found great appeal in educational testing because the requirements could be rapidly implemented, often in relation to elections, and the visibility of the results could demonstrate to the public that elected officials were making progress in the area of school improvement. High-stakes testing was also less expensive than other approaches—like reduced class size—that might impact learning (Linn, 2000).

Other educational researchers made similar claims to Fuhrman (2003) and Linn (2000). For example, Ramirez (1999) wrote that accountability-based reform rested on assumptions that students were unmotivated to learn, teachers lacked necessary skills, local schools were unaware of what students should learn, and that additional testing was necessary to pressure schools toward improvement.

Robinson and Timperly (2000) conducted a qualitative study of 12 New Zealand primary schools to provide an account of the conditions in which educational accountability might prompt student achievement. The researchers concluded that, "Accountability promotes improvement when the accountable agents accept the validity of the judgment made about their performance, accept responsibility for improvement, and have the capacity to achieve it" (p. 68). The researchers also found that the reporting of student achievement was not likely to improve learning, because the teachers understood that parents desired good news, and that any bad news reflected poorly on the teacher. On the other hand, parents were neither in position to interpret nor question the validity of the information. The reports were found to be positively biased, because accuracy threatened both the reputation of the teachers and the esteem of the family.

The idea that accountability alone could improve student achievement had been challenged; moreover, accountability systems provide no prescription for improvement (Hanusek & Raymond, 2001). NCLB put a premium on testing but may not have had any real effect on student learning. Harvey (2003) asserted that it is "easier to talk about test scores than to talk about learning" (p. 20). Guskey and Kifer (1990) concurred. They conducted a quantitative study of the school rank on the Kentucky Essential Skills Test (KEST). The purpose of the study was to determine if correcting for demographic factors provided more meaningful interpretations of test results. The researchers concluded that, "A ranking procedure always produces a top and a bottom, without knowing what either really means. A bottom-ranked school district may be doing quite well on certain criteria and a top-ranked district may be doing quite poorly" (p. 16).

Despite the flaws associated with accountability, some researchers contended that holding schools accountable would bring needed reform. Hess (2003) said that schools had been "constantly reforming without ever really changing" (p. 24), and drew comparisons to the improvement of the auto industry in the late 1970s due to foreign competition "shocking them into action" (p. 24). Additionally, teachers had great latitude in curriculum and instructional practices, resulting in variations of subject matter from one classroom to the next (Jerald, 2003). Schools, and particularly teachers, answered to few outside sources concerning the effectiveness of their instruction and delivery of the curriculum. Hess (2003) asserted that two forms of accountability existed. He wrote that while "nice" accountability assumed the key to school improvement was to provide educators with more resources, "mean" accountability describes coercive measures, and that "school performance no longer relies on fond wishes and good intentions" (p. 23).

Validity. The question arose as to whether any of the accountability systems provided valid methods and measures of schools to facilitate improvement. Researchers continued to study the probability of whether *all* schools could bring *all* kids to required level of proficiency. Under the new accountability systems, output (student achievement) was the measure of school effectiveness as opposed to input (per pupil spending).

Hallows (1988) conducted an economic analysis measuring output performance with data collected during 1979-1980. She found that output levels were significantly influenced by the student characteristics of a district, specifically, the prior levels of student achievement, time spent in school, and the percentage of students planning to go to college. The study also indicated that districts with higher levels of per student spending tended to have a relatively higher level of positive student characteristics. The districts with these characteristics also had a higher probability of achieving a state-normed level of output. Conversely, districts with a lower level of per student spending had lower positive student characteristics and the probability of achieving a state-normed level of performance was lower.

Hallows (1988) also found that the probability of improvement over time favored districts with lower positive student outcomes. Hallows contended that basing state aid grants on increased performance levels would be a more equitable policy. A state aid formula weighing increased performance would benefit poorer districts, due to the fact that those very districts could demonstrate the greatest gains, as opposed to wealthier districts that may be experiencing the ceiling effects associated with high student performance.

A criticism of assessment driven reform was that schools were held accountable in areas they could not control (Ramirez, 1999). Assessments took into account factors not associated with the specific school being measured for accountability. Steffy and English (1997) stated that 89% of the variance on the National Assessment of Educational Progress (NAEP) could be attributed to number of parents living in the home, education of the parents, community type, and the state poverty rate. Fuhrman (2003) referred to holding students to a common level of performance as a status score that "reflects the students' background as much as it does any learning that took place in the year of testing" (p. 3). Harvey (2003) was also critical of the requirements of NCLB when concluding that the federal accountability model will "run up against the universal challenges of race, class and poverty" (p. 18).

NCLB (2002) required that all students and subpopulations be proficient in reading and math by the spring of 2014. The law challenged schools to improve achievement in yearly increments for all students to reach this proficiency standard. Linn (2000) noted that while Texas reported 92% of the students as proficient in 2001, the state averaged a 1% increase in meeting the standards. The AYP requirement of equal incremental increases in student performance lacked empirical basis, was viewed as arbitrary, and was missing a basis in sound theory about school improvement (Elmore, 2003). The idea of holding schools accountable for student performance based on state mandated levels of performance had severe limitations. According to Doran (2003), high stakes tests, "fail to take into account growth in proficiency and unfairly punishes schools serving disadvantaged students" (p. 57). Doran echoed the assertion by Hallows (1988) from 15 years earlier. Hallows proclaimed that schools serving the most needy students had the greatest potential for growth, but were punished by legislation.

Value added accountability. Many critics of state-mandated performance levels presented the concept of measuring schools based on improved performance. Hallows (1988) concluded that schools with high positive student factors had a higher probability of meeting state-normed levels, while schools with low positive factors had greater probability of increased achievement over time. In line with this notion of measuring improvement as opposed to status, Doran (2003) proposed that schools be measured on the value they add to a student's learning. This analysis assumed that both school (i.e., effective instruction) and non-school (i.e., income level) factors have an impact on student learning, so the measurement of increased learning can attempt to control for the non-school factors (Fuhrman, 2003). Linn (2000) suggested that "value added" should be considered, as it allowed disadvantaged schools an opportunity to show improvement, and that comparisons should occur on an annual basis within the school, rather than between schools.

The accountability movement prescribed under NCLB favored high standards for all students. The question arose as to whether high standards for all students equated to having common standards for all students. Linn (2000) suggested that accountability systems did not need common standards to have high standards. Students start with a variety of abilities, and to measure performance solely based on a "one size fits all" mentality negated achievement gains. The measurement of student improvement may prove to be a more realistic method of school accountability.

Public perceptions of schools. The executive summary of the 35th Annual Phi Delta Kappa/Gallup Poll of the Public's Attitude Toward the Public Schools (Rose & Gallup, 2003) shed some light on public perceptions of accountability. Data from the survey showed that the public had high regard for the public schools, was uninformed about NCLB, and that decisions about schools should be left to the state or local district.

Johnson (2003) discussed public attitude toward accountability, citing a variety of public surveys. According to these studies, public school parents believed that standards and assessment help student learning, and the public reporting of test scores would make principals and teachers work harder for improvement. The public also believed it was wrong to base promotion on the results of a single test. It appeared that the public desires schools to be accountable, while the rationale and sanctions imposed by the system are open to debate.

Consequences. One of the facets of the current accountability system was the assignment of sanctions and rewards based on student achievement. Marzano and Kendall (1998) emphasized that, "Policy makers and others are using large-scale assessments to decide whether teachers and schools are providing an adequate education to all students and attaching consequences, positive and negative, on the basis of student assessment results" (p. vii). Sirotnik (1999) contended that the purpose of accountability systems was to use data to place blame, and that the appropriate use of accountability was to provide resources in areas of need rather than serve as a punitive measure. Fuhrman (2003) took Sirotnik's (1999) claim a step further, contending that the assignment of rewards or sanctions assumed that those held responsible, specifically teachers and students, had the ability to make improvements.

NCLB (2002) provided sanctions to schools failing to make AYP for the total population or any subpopulation. If a school failed to make AYP two years in a row, the school was identified for needing improvement. If a school identified for improvement failed to make AYP, the school was selected for corrective action, which included offering public school choice and providing supplemental student services. If a school identified for corrective action failed to make AYP for two years in a row, the local education agency (LEA) had to prepare a restructuring plan.

Sirotnik (1999) and Baker, Linn, Herman, and Koretz (2002) agreed that accountability systems could be based on a single measure of performance. A criticism of the new accountability system was that consequences were often tied to the performance on a single measure as opposed to multiple measures (Fuhrman, 2003). Linn (2000) suggested that it was appropriate to use multiple measures of student achievement rather than putting the weight on a single test.

While the call for school accountability was noble, there remained dissention as to the consequences of the employed system. The debate between those seeing NCLB as well intended, yet misguided, and those that believed sanctions and coercion would force schools to improve would continue as it had vast political ramifications. Only time would tell as to whether the latest attempt at school reform would have a lasting effect.

Assessment

NCLB (2002), the federal accountability model for schools, was predicated on assessment of students to measure student learning and school effectiveness. NCLB required that each state provide public reporting of information about student achievement on state academic assessments, as well as provide a comparison between the achievement levels of each student and the state's objectives. Indeed, the provisions of NCLB and the entire modern accountability age used "assessment" as the vehicle for gauging student achievement and school success.

Assessment Terminology

Assessment was an essential element of the learning process involving school stakeholders such as students, parents, teachers, administrators, and policymakers. Various authors have referred to assessment as systematic processes used to make inferences about student learning based upon multiple sources of evidence. It was this global process of collecting, analyzing and synthesizing information about students that allowed us to understand and describe students and their achievement more effectively (Brown, 1983; McTighe & Ferrara, 1998).

Assessment derived from the Latin root assidere, which means to "sit beside" (McTighe & Ferrara, 1998). Although this may have been a historically accurate depiction, contemporary educators oftentimes reduced the assessment process to a grade, quartile, percentile or raw score. Other widely-used terms in assessment included testing, evaluation, summative assessment, and formative assessment. For clarity in purpose, McTighe and Ferrara's (1998) definitions were used to form the foundation of assessment knowledge. Assessment was gathering and interpreting information in order to construe knowledge, skills, and characteristics of people. Testing was a form of assessment often found in a paper-pencil format, which also frequently involved time limits, restricted access to resources, and a limited range of acceptable responses. Evaluation referred to the process of judging quality, value, or worth correlated to a set of criteria (McTighe & Ferrara, 1998). Summative assessments were described as the degree of knowledge or proficiency attained at the culmination of a program of study. Conversely, formative assessment was more of an on-going diagnosis providing information to assist educators when adjusting instruction and enhancing student achievement (McTighe & Ferrara, 1998; Stiggins, 2002).

Another term was evolving in the assessment arena—alternative assessment. McTighe and Ferrara (1998) noted, "Generally, alternative assessment was used to refer to those assessments that differ from the multiple-choice, timed, one-shot approaches that characterized most standardized and some classroom assessments" (p. 3). Volumes have been written about the topic of alternative assessment and its reliability, or lack thereof, in the assessment and accountability era.

Other assessment terms noted in the literature included standardized, high-stakes, norm-referenced tests (NRT) and criterion-referenced tests (CRT). Standardized tests involved the process of "standardizing" the time, resources, and instructions to provide increased reliability during an assessment session. Although standardized tests were often used for this "one shot" purpose, it was important to note that standardized tests were only considered "high stakes" if the results from such tests were used in making decisions regarding rewards and sanctions for stakeholders (Popham, 2002; Smith,1993). Not all large-scale standardized tests were norm-referenced. Some educators began writing standardized criterion-referenced assessments like the NAEP and some state-level competency tests. In this model, the assessment was measuring students against a set of pre-established criteria in a universal manner regarding the amount of time allotted, the directions provided, and the format of the assessment (Popham, 2002).

Many would suggest that NRTs and CRTs were created with different purposes in mind. However, Harsh (2000) would argue that they, "Direct attention to different uses and references for information and decision-making. Their combined contributions allow a more detailed and comprehensive means of assessing the outcomes of an educational program" (p. 1).

Others described NRTs and CRTs as national tests and state-specific tests. Five national standardized tests were used extensively in the nation's public schools. State specific tests were often constructed like the national standardized tests and done so by the same three major measurement companies (Popham, 2003; Stiggins, 1991). Consequently, Popham (2002) stated, "State-specific tests often perform measurement tasks that were essentially identical to those performed by national tests [norm-referenced tests]" (p. 19). Popham (2002) also clarified the differences between achievement tests and aptitude tests. Achievement tests were intended to measure students' knowledge and skills, whereas aptitude tests, like the Scholastic Aptitude Test (SAT) and American College Test (ACT), served to predict how well a student would perform in subsequent academic settings.

Assessment History

The history of achievement tests can be traced back to World War I, during selection of officers for the US Army (Popham, 2002; Stiggins 2002). The Army Alpha was created and given to almost 2 million Army recruits for the purpose of sorting the examinees based upon relative mental abilities (Brown, 1983; Popham, 2002). According to Popham (2002), the Alpha worked well and became the model for subsequent educational tests built in the United States. Popham stated that, "Today's standardized educational achievement tests are patterned after the Alpha and attempt to carry out its measurement mission of providing relative score-based comparisons of examinees" (p. 19).

Stiggins (1991) reemphasized this point when he wrote that other forces came to usher in a new view of educational assessment near the end of the 1920s. Compulsory school attendance laws and an emerging industrial society contributed to America finding itself needing to educate its large and ethnically diverse population while utilizing the maximum in efficiencies. Stiggins (1991) alleged that in order to sort students and begin the process of determining their futures, "we evolved into an assembly-line method of organizing schools" (p. 265). The assembly line evolved into a progression of grades, adding new information at each of the grade levels. Consequently, schools decided that the assessment of student achievement needed an achievement continuum to determine what students attained within an allotted amount of time (Stiggins, 1991).

The next wave of assessment occurred in the 1960s, when legislators and educators tried to develop a cost-benefit accounting of education so state governments could defend and direct tax revenues for public schools. Due to the difficulty of assessing the effects of education on employment, it seemed—from an economic standpoint—that measuring short-term educational benefits would be a practical audit for education. Thereafter, the skills and knowledge acquired by public school students followed this short-term method (Bock, Mislevy, & Woodson, 1982; McTighe & Ferrara, 1998). Additionally, survey sampling gained momentum in education, and consequently allowed for grouping students rather than assessing individual students to estimate attainment levels in subject matter areas. The survey sampling approach saved time. Lord and Novick's (1968) work was among the first to study the quantitative sampling of items by using subsets of items from domains, known as matrix sampling.

According to Bock et al. (1982), Ralph Tyler led the matrix sampling concept, which generated momentum for a national assessment. Tyler (1973) expanded his work to involve procedures like defining objectives, writing, reviewing and selecting items, and instrument construction and administration. In addition, the concept of random-item reporting became a mainstay for state and national assessments. The work of Tyler and others shaped educational assessment to "estimate attainment in culturally or educationally diverse groups with results detailed enough to be useful, but at costs low enough to be practical" (Bock et al., 1982, p. 11).

The role of educational assessment expanded dramatically in the 1970s and 1980s. Expansion occurred in large part because of reports like the aforementioned *A Nation At Risk* (National Commission on Excellence in Education, 1983) that documented the failure of the American public school system (Marcoulides & Heck, 1994). Bruno and Marcoulides (1985) proclaimed that assessment became the bonanza of education due to the negative publicity. "Assessment instantly became both the means and the end. Tests were the sole measure of educational quality, often leading to short-term remedies at the classroom level" (p. 334). The federal government played a prominent role in the new wave of testing. Assessment became connected to federal initiatives, federal funding, and programs that gauged student learning, particularly of disadvantaged youth.

Research suggested that the assessment programs of the 1970s and 1980s had a negative impact on students and teachers (American Educational Research Association, 1991; Madaus & Kellaghan, 1993; Marcoulides & Bruno, 1986; Marcoulides & Heck, 1988). Although some would contend large-scale assessment systems met the mark, others would caution that the so-called "successes" to date were mostly due to advances in survey methodology, notably the multiple-matrix sampling theory. Because of this theory, it was possible to obtain precise and sufficient estimates of attainment over a broad range of skills with minimal educational resources expended. Bock et al. (1982) contended, "The remaining challenges of assessment, improving methods of reporting results and monitoring them over time, demonstrates the need for commensurate progress in measurement methodology" (p. 11).

Recent assessment history was predicated on the reauthorization of the federal Elementary and Secondary Education Act in 2002, which mandated massive increases in state assessments. Additionally, high-stakes ramifications (sanctions, funding, and public school reputations) were attached to the tests. As schools complied with the federal regulations, state and local testing programs, as well as classroom practices, faced many disruptions (Neill, 2003). Other researchers also noted changes to assessment practices. Buckendahl, Impara, and Plake (2002) explained, "In response to external pressures, control over methods of accountability had shifted in many instances from the local jurisdiction (school districts) to the state jurisdiction (departments of education and legislative agencies)" (p. 6).

Certain experts believed this legislation would permanently harm public education. Neill (2003) shared his view when stating, "It would be worth enduring these difficulties if we could be reasonably sure that the test-driven changes would produce improved learning opportunities and outcomes, particularly for the low-income students. Unfortunately, evidence and reason argue they will not" (p. 18). In the 33rd Phi Delta Kappa/Gallup Poll, 75% of the respondents believed the best way to measure student achievement was to use classroom work and homework (Rose & Gallup, 2001). The recent research suggested that the public was willing to consider multiple sources of information when pondering the accountability of public schools, and would give less credence to standardized NRTs as evidence for achievement. "The major justification given for this push is that the multiple-choice tests of the past two decades were faulty and unable to measure higher order knowledge and skills" (Marcoulides & Heck, 1994, p. 334).

Principles of Effective Assessment

The world of education was still searching for the best way to accurately assess student achievement. Various sources noted that powerful, teacher-based assessment occurred when quality conditions were met (McTighe, 2001; Neill, 2003; Popham, 2003; Roschewski, 2002; Stiggins, 2002). When teachers engaged in the process of systematically reviewing student work using multiple methods, across diverse contexts, and over time, the process of informed assessment occurred (Wolf, 1993). Teacher training to develop assessment literacy was paramount for informed evaluation. First, teachers must possess baseline knowledge about a variety of topics such as curriculum, instruction, childhood development, and diversity roles. This educational foundation must be combined with extensive knowledge of quality assessment methods that meet the standards for reliability and validity (Stiggins, 1995). Wolf (1993) agreed, finding that, "A knowledgeable teacher is the foundation of informed assessment" (p. 519). Multiple assessment methods provided an array of lenses by which to view student work. Blending student interviews, collections of student work, performance assessments, and pencil/paper tests were valuable strategies in evaluation of student achievement (McTighe & Ferrara, 1998; Wolf, 1993). Wolf (1993) suggested that the next steps in assessment literacy for the nation's teachers were greater pre-service and in-service preparation in classroom assessment. Wolf also believed that teachers needed administrative support because, "extending the work of teachers who are practicing informed assessment will require greater respect and commitment from those outside the classroom" (p. 522).

Effective assessment demanded a setting in which teachers worked collaboratively to establish sound and shared visions of curriculum, clear instructional goals, and comprehensive school-wide systems for evaluating and reporting student performance (Wolf, 1993). Lewis (1996) affirmed Wolf's claims when she concluded that teachers were not automatically good writers of tests. She suggested that teachers be the center of assessment activities—embedding assessments in their instruction, scoring the assessments, and discussing standards for good student work with colleagues, parents, and students. Otherwise, poorly developed and implemented standards and assessments would likely become a distraction and source of frustration (Gandal & Vranek, 2001).

Realizing that a few multiple-choice items were only one of many ways to assess student achievement, performance assessments were gaining momentum in many states. Mills (1996) explained that, "open-ended alternative assessment formats, such as performance-based assessment and portfolio assessment, evaluate students more equitably and ultimately improve student learning in the classroom" (p. 2344).

Popham (2002) concurred with Mills, contending that incompatible missions existed when using NRTs to evaluate schools. He proclaimed that, "Our chief concern should be determining the quantity and quality of what students have learned" (p. 19). When decision-makers rely too heavily upon tests with the purpose of providing relative comparisons, the concept of determining what students have actually learned takes a backseat to the comparative data. According to Popham (2003), the reason for this problem was the necessary score-spread, which "leads to the creation of standardized achievement tests that do a dismal job of measuring how much students have learned in school" (p. 19).

Madaus (1988) summarized arguments against top-down types of measurement systems as those focusing only on the skills tested, often of lower levels of thinking, and leaving out other important skills in the area of creativity. Moreover, McTighe and Ferrara (1998) contended that, "large-scale standardized tests typically do not provide sufficiently detailed or timely information regarding student achievement of specific curriculum goals" (p. 5).

Without effective classroom uses for assessment, it was virtually impossible for teachers to know whether or not students were learning what was important for them to learn. McTighe and Ferrara (1998) explained classroom assessment purposes well when writing, "Fairness in classroom assessment refers to giving all students an equal chance to show what they know and can do" (p. 8).

School-based Teacher-led Assessment Reporting System (STARS)

Accountability chose assessment as its vehicle to evaluate the effectiveness of schools. With that in mind, the assessment systems employed to evaluate school effectiveness must be scrutinized to determine overall effectiveness.

Philosophy

One state attempted to keep federal and statewide accountability closer to the classroom. Nebraska's STARS was grounded in the philosophy that assessment must be an accurate reflection of student achievement while maintaining the essential implications for learning at the classroom level. This concept coincided with McTighe and Ferrara's (1998) suggestion that, "The primary purpose of classroom assessment is to inform teaching and improve learning, not to sort and select students or to justify a grade" (p. 1).

Nebraska's "different" system combined student performance, technical quality, and non-cognitive indicators of performance. The Nebraska guidelines, as developed by NDE, required that each public school district adopt measurable quality academic content standards in the core subjects and report results of local assessments to NDE. The STARS guidelines also mandated local assessments of reading, speaking, and listening, participation in a statewide writing assessment, and submission of local assessment procedures to NDE, reviewed and rated by independent assessment experts (Gallagher, 2001).

Pre-empted by state legislation in 2000, which established the requirements and general procedures for implementing standards, assessment and accountability, local districts had to create assessment systems that measured students' performance on state standards within language arts, math, science and social studies, in addition to a statewide writing exam. In Nebraska, "Performance is reported as the percentage of all students meeting or exceeding individual content standards or a collective set of standards" (Buckendahl et al., 2002, p. 8). In the STARS model, local districts determined strategies to measure student performance of the core content areas. These strategies may have included norm-referenced, criterion-referenced, or classroom-based measurement strategies (or a combination of these strategies) (Buckendahl et al., 2002).

The STARS model dictated that districts submit local assessment plans each September. The local assessment plan outlined the assessment procedures used in the district to measure content standards assessed that year. Additionally, districts submitted a culminating DAP by June 30. These portfolios contained information about the assessment measures and the assessment procedures used within individual districts, as well as a report of student achievement results on the content standards. This report was sent to the NDE for review by a panel of 16 independent reviewers, assembled from experts within and outside of Nebraska. The reviewers evaluated the portfolios each year from July 1 through September 1. The reviewers provided school districts with feedback and suggestions for improvement on their local assessment process as well as a rating for the quality of the assessment. Upon completion, results were sent to the school districts.

An additional panel, known as the National Advisory Panel, consisted of well-known assessment experts who provided guidance to the entire portfolio review process, assisted with the training of the portfolio reviewers, and determined model assessment practices within the state (Gallagher, 2001). These model assessment practices were recommended from all school district sizes and circumstances—large, medium, small, urban, and rural—and were models for replication purposes. In the fall, prior to information being shared with local districts, the National Advisory Panel convened to review the model assessment practices identified by the 16 reviewers. The panel recommended a final selection of the most promising practices for public release. The Advisory Panel selected four models for each of the Six Quality Criteria; NDE then disseminated the information to all Nebraska school districts (Gallagher, 2001).

Districts were required to complete reports that included all of the students assessed (including students with disabilities and students learning the English language). Students not included in each of the reporting forms were to be reported as "Not Assessed/Not Included in Reporting." After student achievement had been reported and calculated at the state level, statewide mastery levels were established for student performance. Again, the Buros Center for Testing facilitated these determination sessions, which included participation by a representative group of Nebraska educators from across the state. These mastery levels were determined in order to provide districts with ratings of student performance. The five classifications of ratings were: exemplary; very good; good; acceptable/needs improvement, and unacceptable. School districts received a student performance rating for each of the corresponding grade levels assessed and reported (Gallagher, 2001).

Quality of STARS

The common thread for all local districts in the design and review of their local assessment systems was known as the Six Quality Criteria. These criteria were found in the Definitions section and included the following:

1. Local assessments reflect state standards—there is alignment between the standards and the assessments, and there is sufficiency of coverage to determine adequacy.
2. Students have the opportunity to learn the content before the assessment—the correlation and timing among curriculum, instruction and assessment.
3. Assessments are free from bias—so as to keep the assessments valid.

4. Assessments are at the appropriate level—readability and expectations are congruent to the level of the learner.
5. Evidence of reliability is demonstrated—objective and subjective measures are used as appropriate for the type of assessment.
6. Mastery levels are appropriate—mastery levels are statistically determined not arbitrarily identified.

These criteria represented sound characteristics of quality assessments as defined by the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). Raters evaluated the district assessment system on these criteria and then determined an overall quality rating ranging from unacceptable (1) to exemplary (5) (Buckendahl et al., 2002).

The technical quality rubric designed by the Buros Center for Testing clarified the necessary characteristics for the classification and demonstrated psychometric soundness of methods used for measuring student performance (Plake & Impara, 2000). A matrix was used to demonstrate the interpretations and combination of technical quality. Each individual criterion was rated as met, met-needs improvement, not met. Each level shared specific qualifications for the rating. Broad comparisons were possible for information at the state level, but emphasis and specific analyses were meant for local districts to monitor student achievement.

Independent Review of STARS

The Nebraska assessment model may have been different, but it met the muster of some assessment experts. McREL (2001), for one, conducted an independent, multi-year study reviewing the STARS program and determined that, "Although the STARS program may be unconventional, it may serve as a national model for a healthy assessment system" (p. 4). McREL went on to state that the STARS program was positively influencing curricula and class instruction and fostering a high level of professional development and assessment literacy among teachers and administrators (McREL, 2001).

Many states used a composite index of district performance considering achievement and non-cognitive measures like socioeconomic status, limited English proficiency, and mobility. Other states rank ordered/rated school or district performance using scale scores based upon test performance without considering non-cognitive indicators (Buckendahl et al., 2002). Researchers were finding that rank ordering schools would not answer the question of absolute performance. Buckendahl et al. (2002) commented, "Small differences in performance could translate to large differences in rank order" (p. 7). Guskey and Kifer (1990) noted this problem when they analyzed the statewide data and found how "relative position" may shift dramatically with small performance changes.

STARS was different than other states in a variety of ways. First, the system used local district level assessments, rather than a large-scale state test. Second, the focus was on the classroom. Next, teachers were responsible for developing assessments. Finally, this system relied upon developing assessment literacy in its teachers (Bandalos, 2004). Yet, Nebraska's STARS program was obviously disadvantaged when making comparison among students, schools, and districts. However, Nebraska's commissioner noted that local decisions regarding curriculum and measurement practices allowed assessment to inform curriculum and instruction in the classroom—a critical element according to writers in the assessment field (Roschewski, 2004).

Assessment systems will likely evolve. Recognizing systems that meet federal and state accountability, yet have the power to provide quality information for the classroom teacher seem to be the best of both worlds.

Impact on Educators

While the Nebraska assessment model found its niche with teacher-led, local assessment practices, most other states elected for one statewide test. Regardless of the assessment procedure used to meet accountability requirements, the educator was the individual swimming in the middle of a modern assessment tidal wave.

Numerous researchers conducted a host of studies in the recent past that depicted the manner in which professional educators, regardless of position, evolved since the inception of the contemporary high-stakes accountability and assessment movement. A huge majority of the latest research devoted to the assessment and accountability movement explored its impact on teachers and teaching, but numerous studies related to educational leaders as well.

Impressions of Teachers

It was fair to surmise that the professionals most impacted on a daily basis by the high stakes accountability and assessment trend were teachers. Literature dealing with the connection between teaching and accountability and assessment was best categorized into the three categories of Autonomy Lost, Professional Development, and The Wildcard: Classroom Practice.

Autonomy Lost

With the inception of high-stakes testing, teachers had no choice but to disregard any feelings of isolationism from the past and open their classroom door to the accountability age in America. This new age left many teachers fraught with emotion, feeling violated as professionals in their field (Buly & Rose, 2001; Cimbricz, 2002).

Teachers felt their sense of professionalism was at stake in the high-stakes world of education. This accountability movement left many educators feeling disharmonized with their role as classroom teacher. Mathison and Freeman (2003) emphasized this point in a year-long study conducted of two New York State elementary schools. Their findings indicated that teachers felt engulfed by the requirements of state testing. Mathison and Freeman (2003) stated that, "Teachers work has come to be defined by the state mandated tests as well as district directives geared to improve state scores" (p. 5).

Gallagher's (2001) year-one findings of the previously mentioned Nebraska STARS model were similar to the findings of Mathison and Freeman (2003). In Gallagher's (2001) Executive Summary, derived from numerous surveys, interviews, and observational research, he summarized that teachers felt "understandably fearful that they were being ‘deprofessionalized’ as their workload intensified and the screws of accountability are tightened” (p. i).

The “deprofessionalized” educators believed the current system compromised best practices in favor of attempting to appease legislators, administrators, parents and state mandates (Brunn, 2003). Teachers were forced into difficult, no-win situations with irresolvable teaching dilemmas. The question of using teacher judgment to meet the needs of one’s particular classroom was challenged by the looming state test. More often than not, teachers succumbed to instruction that will help students perform well on state tests while sacrificing professional decision-making (Buly & Rose, 2001; Mathison & Freeman, 2003). Beran (2003) researched 257 fourth-grade teachers in Nebraska and found that morale was down because of state standards. According to Beran, “The stress caused by Nebraska’s state standards process was noticeable. Teachers see morale declining as the additional workload increases” (p. 52).

A concern for the unknown was another factor creating unrest among professional educators. Grant’s (2000) two-year study of cross-sectional teachers in New York State found an uneasy combination of anticipation and dread regarding a revision to the state test. Grant suggested that, “Teachers ideas and voices were largely ignored as those above them-state and district-level actors-did the real work of policy change” (p. 9).

Gandal and Vranek (2001) built upon Grant’s research regarding the lack of teacher input and notification in the accountability and assessment movement when he cited a January 2001, *Education Week* national survey that probed U.S. teachers’ opinions of high-stakes testing. The report declared that, “Only four states let teachers know how each student performed on every multiple-choice item, and only nine states send teachers their own students' scored work on essay questions" (p. 12). Therefore, teachers' faith in tests and testing as a means of gauging student achievement may have proven to be a façade if teachers were ignored throughout the policy framing and reporting process (Boss et al., in press; Brunn, 2003; Grant, 2000).

This new challenge in the profession of teaching created a heightened level of anxiety for many teachers. The possibility of additional stress on teachers was one of Yung's (2001) findings after extended classroom observations and interviews of ten Hong Kong teachers becoming accustomed to a new high-stakes assessment model in their country. Yung suggested that teachers who conducted high-stakes assessments for the purpose of meeting a mandate rather than gauging student achievement were likely to feel additional pressure in their profession and not form the connection between the role of teacher with that of assessor. Research conducted by Brunn (2003) concurred with Yung's (2001) conclusions about further stress on the teaching profession. His case study of seven elementary teachers in Colorado found that individual teachers handled the additional demands of state testing differently, and added that both teacher and student anxiety were apparent in the classroom during the days prior to the state assessments.

Another area where many educators felt teacher autonomy had been compromised to the accountability movement was in relation to classroom time and direction of instruction. The implementation of mandated testing forced adjustments in curriculum and instruction to accommodate each state's testing regimen. The first classroom constraint was related to necessary preparatory opportunities prior to the state test.

Amrein and Berliner (2002) examined 18 states with high stakes testing and concluded that a considerable amount of instructional time was spent preparing for the test. Amrein and Berliner (2002) stated that, "teachers believe they spend an inordinate amount of time on drills leading to the memorization of facts rather than spending time on problem solving and the development of critical and analytical thinking skills" (p. 9).

Mathison and Freeman's (2003) research supported Amrein and Berliner's conclusions and pointed to the paradox between state mandated testing calendars with individual student achievement. Her year-long study of two New York State elementary schools found that teachers walked the tightrope of upcoming testing dates with that of helping students understand the material. Re-teaching and individual classroom dynamics become subservient in a high stakes testing atmosphere—teachers must push forward despite the natural nuances involved with a classroom of children.

Research to support Mathison and Freeman's (2003) notion was found in a study conducted by Kaiser and Lambdin (1996). These two researchers studied the Connected Mathematics Project (CMP), an innovative approach to teaching mathematics, but not a high-stakes mandated test. After conducting a qualitative study of 44 middle school teachers across the nation involved in field testing the CMP, Kaiser and Lambdin concluded that, "There are numerous time issues associated with teaching in the spirit of the current mathematics education reform movement" (p. 29).

Professional Development

For teachers to be effective, especially those in states that used performance-based assessments, assessment literacy was a key element that must be developed. A quantitative study of 893 teachers in 34 schools by Bol, Stephenson, O'Connell, and Nunnery (1998) suggested that teachers were poorly equipped to determine the effectiveness and quality of assessments. In fact, these teachers were actually demanding lower order processes such as recognition of details and facts rather than the higher order thinking skills they desired. Indeed, the accountability and assessment movement was predicated on teachers like those in Bol et al.'s study becoming more adept at assessment literacy. Without an improvement of teachers' knowledge-base, the movement will be one of following standards and preparing for state mandated tests, but not addressing student achievement (Bol et al., 1998; Kahn, 2000).

For the teacher to make good, sound rational decisions about education, an understanding of the goal for testing must be understood (Bol et al., 1998; Daniel & King, 1998; Popham, 2003). Daniel and King's (1998) quantitative research of 95 elementary and secondary teachers from two schools in southern Mississippi punctuated the need for teacher assessment literacy to impact achievement. Daniel and King found that, "Teachers recognized that their ability to make an interpretation of data that assesses the student against established standards was a necessary component of teacher effectiveness" (p. 342). Based on this research, Daniel recommended that inservice training be part of a continual study of assessment in the professional career of teachers.

A second reason for professional development related to building teacher ownership in the accountability and assessment process. Falk and Ort (1998) studied 250 New York State teachers, using three sources of information (questionnaire, selected interviews, and observation), and revealed that "Teacher involvement in assessment—

both use and scoring—enhanced and supported their learning and strengthened their sense of professionalism, and built their support for this change" (p. 61). The involvement Falk and Ort observed in professional development training created an environment for teachers to "learn from one another and to validate their knowledge as competent professionals" (p. 62).

The findings of Falk and Ort (1998) were complimented by research of the STARS model in Nebraska. This aforementioned system built assessments at the classroom level, creating a great deal more empowerment of teachers. Bandalos' (2004) qualitative study of Language Arts teachers in Nebraska's first year of implementation of locally developed assessments found "greater teacher use of results for instructional planning, greater diagnostic utility of the assessments, and gains in assessment literacy" (p. 33).

Teacher collegiality was another ingredient for professional development in today's high-stakes testing arena. Grant (2000) pointed toward collegiality as a means of professional development. His aforementioned study found that with teachers, "The power of such informal relationships was apparent: These teachers sense they were working with peers who held similar goals and concerns, who were willing to share ideas and practices, and who offered a sense of belonging" (p. 8).

Professional development for those "in the trenches" developed a sense of confidence as well. They developed relationships in such a manner as to work in the fashion of a team and see the mandated world of assessment through a different lens. This sense of belonging increased assessment literacy, built ownership in the classroom and the school, and promoted an aura of collegiality for the professional educators (Boss et al., in press; Daniel & King, 1998; Grant, 2000). In their case study of four exemplary schools in Kentucky, Wolf, Borko, Elliott, and McIver (2000) described this process of professional development as building "human capital" (p. 388); the creation of a more refined educator who was dedicated to curriculum, instruction, and assessment (Wolf et al., 2000).

The Wildcard: Classroom Practice

The third theme gleaned from the literature was labeled "The Wildcard: Classroom Practice," because teachers were like other human beings in that they had different likes and dislikes, different strengths and weaknesses, and, therefore, different opinions on what to teach and how to teach it.

Expert researcher Michael Fullan (1991) mentioned the role of the individual within a system undergoing reform in his book, *The Meaning of Educational Change.* Fullan pointed to two types of reform, "restructuring" and "reculturing." Restructuring was the sort of reform that provoked a change of the administrative process, but did not affect the foundation of classroom teaching and learning in the schools where it had been introduced. The second concept, reculturing, was a change in the total environment of the organization that created a different way for people to perform their job. Therefore, reculturing massaged the unique traits of individual teachers in such a manner that it impacted the teaching and learning core of the school.

Flett and Wallace's (2002) interpretive case study, based largely on interviews and observations of three middle school science teachers implementing change in

Victoria, Australia, revealed Fullan's (2001) "restructuring" theme in the classroom practice of teachers. Flett and Wallace (2002) reported that this Australian school's changes were "cosmetic" (p. 310) in nature and had little impact on actual classroom instruction. Restructuring of the curriculum occurred, but little change was evident in terms of the order in which lessons were taught and the length of time spent on particular sections of the course by each teacher. Flett and Wallace (2002) surmised from this research that, "Mandated curriculum reform at the classroom level was difficult to affect due to the considerable autonomy afforded teachers" (p. 330).

The "Wildcard" of classroom practice was also evident in Wirszyla's (2002) study of eight physical education teachers in three South Carolina schools implementing mandated change. This study confirmed Flett and Wallace's (2002) suggestion that mandated curriculum reform at the classroom level was difficult to affect because of the considerable degree of independence afforded teachers and the focus of schools toward structural changes. Wirszyla (2002) found that the physical education project's success or failure ultimately resided with the classroom teacher. In her study, "The lead teachers emerged from the data as the integral facilitator to a successful physical education program" (p. 10).

Wirszyla (2002) was careful to emphasize the importance of administrative support in enacting mandated change by stating that, "Teachers are the main vehicle for change, but they cannot be left on their own to develop good programs. They need mentoring, support and accountability" (p. 16). Wolf et al. (2000) affirmed Wirszyla's (2002) claim regarding the administrative role afforded teachers in mandated change while punctuated the importance of classroom teachers in the reform effort. Her case study of four exemplary schools in Kentucky argued that, "The teachers' responses to large-scale reform efforts exist in a larger web of connection and are dependent on their collaborative and consistently positive stance toward learning as well as their principal's leadership" (p. 349). Therefore, administrative leadership and support were positive factors in Fullan's (2001) reculturing process, but pure reform would only occur if classroom teachers were willing to change their everyday instructional practices (Flett & Wallace, 2002; Wirszyla, 2002; Wolf et al., 2000).

The classroom teacher's level of autonomy, their growth in professional development, and their everyday instructional decisions in the classroom were an important dynamic in today's world of high-stakes mandated testing. The individual teacher's decision to develop as a professional or not, to understand assessment or not, to follow the curriculum or not, to teach to the test or not, or simply follow their own path plays a major role in the assessment and accountability movement. In the end, Cimbricz (2002) summarized the impact of teachers on the assessment movement when she wrote that the "Influence of state mandated testing depends on how teachers interpret the testing and use it to guide their actions" (p. 1).

Educational Leadership

Educational leaders underwent the change process in the same world of accountability as teachers. There existed significantly more research connecting teachers with accountability and assessment than educational leaders; therefore, for this portion of the literature review the term "educational leaders" pertains to all of those educators holding leadership positions in the public school arena.

The leadership role in public schools was assuming a new dimension caused by national and international forces influencing our educational system. Long past was the time when educational leaders concentrated on management of an organization. The educational leaders wear many hats, with assessment and accountability being only two of the many pieces of headwear assigned to the leader. Regardless of changing job descriptions or increased workload, educational leadership was a key component of current school systems (Burns, 1998; Fullan, 2001; Cimbricz, 2002; Flett & Wallace, 2002; Wirszyla, 2002).

Waters, Marzano, and McNulty (2003) conducted a study focusing on modern educational leadership. They found that effective leadership impacted the entire school; effective leadership added value to the impact of classroom and teacher practices and ensured that lasting change flourished. Miller's (2003) research on teacher, school, and leadership practices reiterated the emphasis on educational leadership in schools. Miller found that strong educational leadership added inherent value to the impact of classroom and teacher practices, and increased the potential that lasting change flourished.

Other research agreed that educational leaders held the potential to have major impact on student achievement. Marzano (2003), for one, concluded that, "Leadership is a necessary condition for effective reform relative to the school-level, teacher-level, and the student level" (p. 172). Lambert (2003) agreed that educational leaders of today were overburdened, but that they must emphasize student achievement. She concluded that school decision-makers must, "model certain leadership behaviors; abide by certain structures, processes, and policies, and focus on student learning" (p. 80). Indeed, the modern educational leader was under enormous pressure due to increased external accountability requirements, but must not waver from holding student achievement as a top priority.

"Accountable" Leadership

The educational setting in many states was "high stakes." High stakes can be interpreted to mean serious consequences for low scores. Amrein and Berliner (2002) researched high stakes testing and found that punishments were attached to school scores twice as often as rewards. They concluded that 45 states held schools accountable for test scores by publishing district report cards or using rating systems. Their research also determined that, "Fourteen states have the power to close, reconstitute, or take over low performing schools; 16 have the authority to replace teachers or administrators; and 11 have the authority to revoke a school's accreditation" (p. 3). Indeed, the term "high stakes" was an appropriate term for the new age of accountability in education.

Certain studies investigated how schools and educational leaders should react to the new high stakes atmosphere surrounding public education. One study of educational change in elementary schools during the assessment and accountability era by Strahan et al. (2003) found that a shared stance toward learning and decision-making formed links of shared values and beliefs that led to a positive attitude for everyone. The tension associated with high stakes testing was minimized by group ownership of the testing demands.

The research of Marzano (2003) supported the previous study in that he found teamwork and unity were important elements to success in a new era of education. Marzano identified three principles associated with strong leadership in contemporary public education. Marzano concluded that small groups should lead the change process, leadership teams should be decision-making bodies while showing respect for the large group, and that interpersonal relationships remain essential to the functioning of a school. Reeves (2004) also addressed the current need for educational leadership in a time of increased accountability. Reeves concluded that, "It is a moral principle of leadership that no teacher or staff member will be more accountable than the leaders in the system" (p. 20).

Carr and Harris (2001) identified three key areas that should be the basis for leadership decisions in a climate of assessment—resources, programs and practices, and results. Carr and Harris asked educational leaders to ask themselves, "Where should dollars, time, and people be invested to maximize impact on student results?" (p. 63). The Vermont Department of Education (1999) provided educational leaders similar advice. The suggestion for school decision-makers was to focus on student performance, examine local, state, and national performance results, and then set data-driven goals.

Kentucky was one state that took the lead in high stakes testing. Wolf et al. (2000), studied four exemplary school districts as they worked to meet the demands of the Kentucky Education Reform Act (KERA). The research team concluded that, "Leaders worth following . . . work to distribute the leadership rather than guard it for themselves" (p. 386). Certainly, strong educational leadership was critical in an era of high stakes testing.

Leadership Perceptions

Research pointed to disenchantment among educational leaders regarding accountability. One prominent piece of research was that of Milne (2000), who surveyed ten percent of Virginia principals to determine the perceptions of educational reform in Virginia. Results of the study revealed that more than 95% of the principals felt that pressure to improve student performance was unjust, while only 30% of principals agreed that the Virginia assessments gave an accurate description of student achievement.

Burns (1998) studied the perceptions of Oklahoma rural school leaders regarding educational reform in the state. She concluded that rural building-level administrators felt the state mandated reform had little regard to needs of the individual district in the areas of class size, teacher salaries, funding, programs, curriculum, and parent involvement. Similar sentiments were expressed in the findings of a case study done by Schuttloffel (2000) regarding the Kentucky Education Reform Act. Schuttloffel found no magical answer to implementing state educational reforms and raising test scores. She concluded that, "What might prove successful in one school district may confront multiple restrictions within another school, such as an incompatible school culture, a reluctant parent community, and minimal teacher support" (p. 4).

Wolf et al. (2000) studied the same Kentucky model, but concentrated on four of the highest achieving schools in the state. They found different results than Schuttloffel (2000). In fact, they stated that, "Participants wanted to improve the quality of their curriculum, instruction, and assessment. They talked to, with, and about each other, and their tones were consistently tinged with pride and respect" (p. 388).

Wolf et al. (2000) were in a strong minority regarding the attitude of educational leaders about external accountability. Weichel (2002) painted a much different picture of educational leaders and their experiences with state mandated change. In a study regarding the perceptions of Nebraska high school principals, Weichel (2002) concluded that principals believed the state assessment and accountability movement added stress, time, and pressure on educational leaders. Furthermore, the research of Weichel agreed with Schuttloffel (2000) and Burns (1998) that educational leaders were unconvinced that state standards and accountability have had a major impact on student learning.

Summary

The change process was in full gear for educators across the nation. Regardless of the approach each state has taken to meet the guidelines of the modern accountability movement, the prevailing theme in the research field was that numerous challenges existed with the assessment and accountability movement. The STARS approach in Nebraska used a different method to complete this mission. "Teacher-led" was more than a motto in Nebraska. This different assessment system was in its formative stages and worthy of study. Chapter 3 was devoted to the research approach devised to study the Nebraska model.

CHAPTER 3

METHODOLOGY

Introduction

This study was part of a larger four-part study that investigated the problem of implementing a state accountability system through locally developed assessment and reporting of student performance. The four researchers were: Toby Boss, Dan Endorf, Tamara Heflebower, and Phil Warrick. The purpose of this quantitative, descriptive study was to survey and describe the perceptions of Nebraska principals regarding the implementation of STARS. The other parts of the larger study included: a study of Educational Service Unit (ESU) staff developers perceptions by Tamara Heflebower, a study of teacher perceptions by Dan Endorf, and a study of assessment coordinators perceptions by Toby Boss.

The reauthorization of the Elementary and Secondary Education Act of 2002 required each state to submit an accountability plan to the United States Department of Education. Forty-eight states in the country responded to the federal mandate by developing statewide tests. Nebraska was one of two states that chose a different approach. In this state, schools were required to administer locally developed assessments to measure the academic content standards and report the student achievement results to NDE. Since this model was bottom-up approach, the perceptions of educators about STARS was paramount to an analysis of the system (NDE, 2002).

In partnership with Buros, the state department of education developed a set of criteria to determine the quality of the local assessment system. The schools were required to submit a Distance Assessment Portfolio (DAP) which outlined the procedures used to meet the Six Quality Assessment Criteria (NDE, 2002). The district portfolios were evaluated by outside assessment experts to determine the technical quality of the local assessment systems. The student performance and assessment quality results were published as part of the State of the Schools Report. Moreover, state accreditation regulations, NDE Rule 10, mandated schools to comply with the provisions of the state assessment system (NDE, 2004).

In the modern era of educational accountability, the STARS model presented a different approach to assessment. The experience of developing and implementing STARS changed the way many educators in Nebraska conducted business. This new assessment model created interesting challenges for educators at the classroom, district, and state level. To that end, this study concentrated on the perceptions of Nebraska educators as they experienced the development and implementation of STARS.

Research Questions

Three research questions guided the four-part study: The term "educators" encompassed the four different groups being studied. In this particular study the term "educators" represented principals. The three research questions included:

1. What are the perceptions of educators about STARS as it related to education in Nebraska?
2. What are the perceptions of educators about the curriculum, instructional, and assessment practices used to implement STARS in Nebraska?

3. What are the perceptions of educators about the impact of STARS on the professional abilities of educators across the state of Nebraska?

Research Design

Designed as a descriptive quantitative study, this research project specifically analyzed the perceptions of educators involved with the STARS process. Descriptive studies describe a particular phenomenon and are particularly valuable when an area is first examined (McMillan, 2000). Cohen and Manion (1994) wrote that descriptive studies set out to describe and interpret a certain phenomenon. Descriptive research demonstrated the relationship between a preceding event and a present condition (Best, 1970).

Data were collected through a web-based survey developed by the research team. Web-based surveys have the potential of bringing efficiencies to self-administered questionnaires not possible with paper-pencil surveys, all the while reducing the implementation time (Dillman, 2002).

This survey design allowed for a numeric description of the sample by asking educators questions, which then empowered the researcher to generalize to the larger population (Fowler, 1988). This study used a cross-sectional survey procedure to gather data from the sample. Cross-sectional methods seek to gather data from a particular group at a single point in time (Ay, Jacobs, & Razavieh, 2002; McMillan, 2000). This survey gathered data to describe the perceptions and practices of educators who implemented the STARS.

The survey consisted of three sections. The first section gathered data regarding the perceptions of educators about STARS as it related to the improvement of education; the second section concentrated on perceptions of educators regarding the curriculum, instruction, and assessment practices schools implemented as a result of STARS; the final section dealt with perceptions about the professional abilities of educators as a result of STARS. The survey also gathered demographic information to determine that respondents were a representative sample of the population.

Population and Sample

The survey sample consisted of Nebraska public school educators who participated in STARS and held the positions of assessment coordinator, ESU staff developer, principal, and language arts or mathematics teacher in the grades of 4, 8, and/or 11. Since the survey was web-based, collection of e-mail addresses was critical to the success of the sample. The researchers used a variety of sources to develop each e-mail database, including NDE databases, ESU databases, and individual school district contacts. The accuracy and availability of e-mail addresses was a critically important and a limiting factor in the determination of survey participants. Upon completion of the e-mail database, the survey was launched with informed consent communicated in the introduction. Participants had a three-week timeframe to complete the survey during late January and early February, 2005. Following is a description of each educator group and the corresponding sample strategy or population:

Assessment Coordinators

Assessment Coordinators were chosen for the study because they generally facilitated the local STARS process within their school district and guided staff through the technical requirements of the assessment system. Schools designated district staff such as curriculum directors, building principals, and in some cases teachers, to coordinate the assessment process in each district. In many cases, the assessment coordinator facilitated the submission of the DAP, a summary of the technical quality of the locally developed assessment system.

The survey list of assessment coordinators was derived from an NDE database of assessment contacts. The NDE Assessment Office communicated policy updates through the list of assessment contacts, and was the most accurate list of all educators serving as assessment coordinators. While the NDE list was the most accurate, it did not contain e-mail addresses for each contact. These contacts were sent a mailing, which invited them to send an e-mail address if they desired to be included in the list of participants. A total of 386 e-mail invites were sent from the initial list of 500 assessment contacts. The contact list represented each county in the state of Nebraska and each of the six statutory classifications for Nebraska public schools.

Educational Service Unit Staff Developers

ESU staff developers were chosen for the survey because they provided staff development and technical expertise necessary to design and implement quality assessments at the local level. In many cases, staff developers guided curriculum processes, assessment design and refinement, as well as assistance for technologically managing assessment data at the local, district, or ESU level. In addition, ESU staff developers also facilitated district assessment portfolio sessions that helped assessment coordinators organize and write the portfolio. In some cases, ESU professional developers coordinated assessment consortia. The survey list of ESU staff developers occurred through an e-mail database provided by the ESU Professional Development Affiliate Organization and was made available to all 54 staff developers.

K-12 Language Arts and Mathematics Teachers

Teachers were chosen for this study because they either wrote or selected assessments to measure the content standards. Teachers administered assessments as part of classroom instruction, scored and recorded the results. Throughout this time, teachers ensured that methods and procedures met the Six Quality Assessment Criteria.

A stratified, purposeful technique was used to determine the sample. The survey was available to a total of 902 Mathematics and Language Arts teachers from schools in each of the five regions and represented each of the six statutory classifications for Nebraska public schools. At least 175 eligible teachers from each of the five geographic regions in the state received an invitation to participate in the survey. The researchers over-sampled non-Class III districts in order to increase response rates from smaller populations. The regions were based on a geographic clustering of educational service units: Northeast–ESU 1, 2, 7, 8; Southeast–ESU 4, 5, 6; Central–ESU 9, 10, 11, 15. 16. 17; Panhandle–ESU 13, 14; Metro–ESU 3, 18, 19. The sampling technique ensured both a similar number of teachers from each region and a representative sample of teachers according to the six statutory classifications. As previously stated, the informed consent notice was communicated to respondents as the survey was launched. While every effort was made to randomly choose teachers from each classification within a region, the availability of e-mail addresses was an important component in determining the sample. Since no single, accessible database of teacher e-mail addresses existed, the survey list of teachers derived from a combination of ESU, NDE, and individual school sources.

Principals

Principals were chosen for the survey because of their role as teacher evaluator and leader of learning in each building. A stratified, purposeful sampling technique was also used to determine a sample of public school principals for the survey. The survey was made available to at least 60 K-12 principals from the same five geographic regions used in the teacher sampling, for a total sample of 350 K-12 principals statewide. Just like the teacher sample, the sampling technique ensured an equal number of principals from each region, as well as a representative sample of principals according to the six statutory classifications. As with the other surveyed groups, the informed consent notice was communicated to respondents as the survey was launched. While every effort was made to randomly choose principals from each classification within a region, the availability of e-mail addresses proved once again to be a critically important component in determining the sample. Since no single database of principal e-mail addresses existed, the survey list of principals developed from a combination of databases from ESUs, the Nebraska Council of School Administrators, the NDE Directory, and contacts with individual schools.

Survey Instrument and Procedures

The data collection process of this study included information obtained from a web-based survey developed by the four researchers in conjunction with the Nebraska Education And Research (NEAR) Center at the University of Nebraska-Lincoln. The instrument consisted of 59 items and used either a 3, 6, or 7 point Likert Scale to measure the perceptions and practices of respondents. The survey included one open-ended item at the end of the survey that provided an opportunity for respondents to communicate other perceptions about STARS. A copy of the survey instrument is located in Appendix C.

The survey instrument totaled three sections, each correlated to a specific research question. Section 1 of the survey correlated to Research Question 1, regarding the perceptions of educators about STARS as it related to education in Nebraska. This survey section included 10 questions on a 7-point Likert Scale: Much Worse (1); Worse (2); Slightly Worse (3); About the Same (4); Slightly Better (5); Better (6); and Much Better (7). The total score for the section could range between 10 and 70, with a higher score indicating greater agreement that STARS improved education in Nebraska.

Section 2 correlated to Research Question 2 regarding the curriculum, instruction and assessment practices used to implement STARS. Two sub-sections of the survey gathered this data. The first sub-section asked about the importance of sound curriculum and assessment practices due to STARS. Survey questions stemmed from a meta-analysis of curricular practices in *What Works in Schools* (Marzano, 2003), as well as sound assessment practices determined by the Six Quality Assessment Criteria (NDE, 2001). The first survey sub-section included 21 questions on a 6-point Likert Scale: Strongly

Disagree (1); Disagree (2); Slightly Disagree (3); Slightly Agree (4); Agree (5); and Strongly Agree (6). The total score for the section could range between 21 and 126, with a higher score indicating greater agreement that STARS implementation led to more effective curriculum and assessment practices in Nebraska schools.

The second sub-section of Research Question 2 included 12 questions about the importance of instructional practices due to STARS. Survey questions emerged from *Classroom Instruction That Works* (Marzano, 2001), a meta-analysis of classroom instructional practices and a natural connection to the aforementioned curriculum and assessment framework. Respondents ranked the frequency of each instructional strategy on a 3-point Likert Scale: Less Often (1); About the Same (2); and More Often (3). Total score for the section could range between 12 and 36, with a higher score indicating that implementation of effective instructional strategies occurred more often as a result of STARS.

Section 3 aligned to Research Question 3, regarding the perceptions of educators about the professional abilities of other educator groups as a result of STARS. This section of the survey asked questions concerning the degree to which STARS changed the knowledge of educator groups about curriculum, instruction, assessment, and educational leadership among the four groups (assessment coordinators, ESU staff developers, principals, and teachers). This design strategy allowed each of the four survey groups to rate the abilities of the other three groups. The survey section includes 16 questions on a 7-point Likert Scale: Much Worse (1); Worse (2); Slightly Worse (3); About the Same (4); Slightly Better (5); Better (6); and Much Better (7). The total score for the section could range between 16 and 112, with a higher score indicating that the knowledge and skills of various educator groups improved as a result of STARS.

As stated in the sampling section, e-mail communication delivered an invitation to complete the survey. A hotlink within the e-mail took participants directly to the informed consent page found in Appendix C . Participants were required to click on an icon to enter the survey, thereby acknowledging their willingness to complete the instrument and participate in the survey. The electronic survey format allowed for the greeting to include the purpose of the survey, future delivery of communication and information, an avenue to contact the researchers, and an option to opt out of the survey The website was open for data collection during a three-week period of time in late January and early February of 2005. In order to improve the response rate, the researchers launched a second e-mail during the second week that reached only those respondents that had not completed the survey. Such participants were determined through the survey software, thus maintaining complete anonymity. A final e-mail was sent late in the second week, to those respondents that had not completed the survey using the aforementioned method. This message re-stated the purpose of the survey and the upcoming deadline for completion.

Content Validity

Determination of survey content validity occurred through feedback from a panel of eight educators, knowledgeable about the STARS process and ineligible to participate in the study. The panel was asked to evaluate each survey question and provided written feedback about survey questions as well as incidents of bias. The researchers reviewed survey questions that provided evidence that received critical comments from the pilot group and modified the instrument when deemed appropriate.

Reliability

Reliability was defined as the extent to which the findings would be similar if the study were repeated (Ay et al., 2002). It was imperative that survey questions within each subsection measured the same construct, such as the respondent's perceptions of STARS. Coefficient alpha was calculated from the returned surveys to determine internal consistency. Table 1 shows the coefficient alpha statistics of each sub-section for each of the four groups of educators completing the survey instrument.

Table 1

Coefficient Alphas by Educator Groups

Sub Section	Number of Items	Assessment Coordinators	ESU Staff	Teachers	Principals
Perceptions of STARS	10	0.95	0.91	0.92	0.95
Curriculum and Assessment	21	0.95	0.89	0.94	0.97
Instructional Practices	12	0.93	0.96	0.88	0.92
Professional Abilities	16	0.96	0.92	0.97	0.98

Based on the coefficient alpha statistics, the researchers determined that the reliability of each subsection was sufficient for data analysis purposes.

Data Analysis

Descriptive research describes what is factual and accurate. Data analysis portrayed average perceptions of the sample rather than determining causal relationships between variables. The use of descriptive statistics allowed for characteristics of the data to be communicated, and to estimate the characteristics of the population (McMillan, 2000). Survey data were collected and analyzed as they aligned to the research questions. Analysis of data occurred by calculating the percentage of responses in each category, as well as the mean, mode, standard deviation, and variance of each item. The researchers utilized sectional totals to describe the perceptions of the respondents about each research question. From this analysis, the researchers described the prevalent perceptions relating to the particular STARS construct. Simply put, this allowed the researchers to determine if the respondents tended to fall within a particular parameter of the survey instrument, such as “Agree” or “Disagree.”