

5-2012

Perceptions of Kentucky Educators Concerning the Kentucky State Assessment System as an Accurate Reflection of Student Learning

Benny C. Lile

Western Kentucky University, lile@scrctc.com

Follow this and additional works at: <https://digitalcommons.wku.edu/diss>

Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Other Teacher Education and Professional Development Commons](#)

Recommended Citation

Lile, Benny C., "Perceptions of Kentucky Educators Concerning the Kentucky State Assessment System as an Accurate Reflection of Student Learning" (2012). *Dissertations*. Paper 24.
<https://digitalcommons.wku.edu/diss/24>

This Dissertation is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Dissertations by an authorized administrator of TopSCHOLAR®. For more information, please contact topscholar@wku.edu.

PERCEPTIONS OF KENTUCKY EDUCATORS CONCERNING THE KENTUCKY
STATE ASSESSMENT SYSTEM AS AN ACCURATE REFLECTION OF STUDENT
LEARNING

A Dissertation
Presented to
The Faculty of the Educational Leadership Doctoral Program
Western Kentucky University
Bowling Green, Kentucky

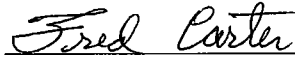
In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Education

By
Benny C. Lile

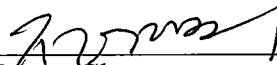
May 2012

PERCEPTIONS OF KENTUCKY EDUCATORS CONCERNING THE KENTUCKY
STATE ASSESSMENT SYSTEM AS AN ACCURATE REFLECTION OF STUDENT
LEARNING

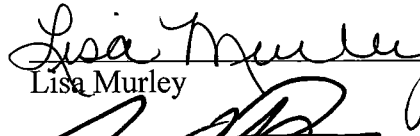
Date Recommended 3-13-2012



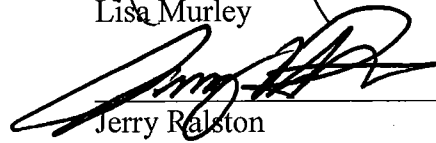
Fred Carter, Director of Dissertation




Kyong Chon



Lisa Murley



Jerry Ralston


Dean, Graduate Studies and Research

16-APRIL-2012
Date

To:

My wife Chris,

and my daughters Cameron and Casey.

Thank You

ACKNOWLEDGMENTS

An effort of this magnitude creates an exhaustive list of those deserving recognition. I would like to start at the beginning. Most of all I give thanks to my Lord and Savior, Jesus Christ, who created all things, including me. I thank my parents, Lavon and Glynna Lile, who showed me by example the meaning of effort, honesty, and love. To my brother Marty and best friend Tom Cravens, who have always shown interest and support.

To my students I taught at North Metcalfe Elementary. You all were, and continue to be, an inspiration to me. I extend a debt of gratitude to my colleagues from Metcalfe County Schools, the Kentucky Department of Education, and Barren County Schools. A very special thank you goes to Vicki Weaver, Shari Alexander, Jeanelle McGuire, Melissa Moss, Melinda Owens, Valerie Stokes, and Dinah Wallace. I would have never finished this without your patience and assistance.

I thank my church family at Coral Hill Baptist, especially my pastor Dr. Ray Woodie, who has been a continual source of encouragement. Ken Draut was an invaluable resource as I worked through this project. He and the OAA staff at KDE have invested unknown time and effort for the students and staff of Kentucky schools. They are much appreciated. The officers and board members of the Kentucky Association for Assessment Coordinators went above and beyond the call in accommodating my research.

Additionally, I thank all my classmates from Cohort II. An extra-special recognition goes to “The Boys,” Adam Murray, Chris Schmidt, Bob Jackson, Tom Stewart, and Tony Kirchner. Just getting to know you all was worth the price and the

effort of this endeavor. Also, I thank Matthew Constant and Wesley Waddle from Cohort I, for guidance and information.

A large debt of gratitude goes to my committee members. To my chair, Dr. Fred Carter, we have both survived my first experience with a dissertation and your first experience as chair, Dr. Kyong Chon for unbelievable patience, Dr. Lisa Murley for assistance throughout the program, and Dr. Jerry Ralston for providing ongoing encouragement throughout the process.

And last, to Chris, Cameron, and Casey. For every event I missed and every piece you all had to pick up when dad wasn't there, thank you. Your unflinching love and support from beginning to end made this all possible.

And though you be done to the death, what then?
If you battled the best you could;
If you played your part in the world of men,
Why the Critic will call it good.
Death comes with a crawl, or comes with a pounce,
And whether he's slow or spry,
It isn't that fact that you're dead that counts,
But only how did you die?

Edmund Vance Cooke

TABLE OF CONTENTS

LIST OF TABLES.....	ix
ABSTRACT.....	x
CHAPTER I: INTRODUCTION.....	1
Problem Statement.....	2
Purpose and Background.....	6
International Comparisons.....	8
Finland.....	8
Singapore.....	8
Costa Rica.....	9
Research Questions and Hypotheses.....	11
Significance of Study.....	12
Definition of Terms.....	13
CHAPTER II: REVIEW OF THE LITERATURE.....	16
Introduction.....	16
Effects on Results.....	19
Sub-population Issues.....	22
State Issues.....	24
Kentucky and Maine.....	25
Maryland.....	26
Colorado.....	26

Various States.....	27
Functions of Test Instruments.....	28
Effects on Schools and Instruction.....	32
Instructional Practices.....	35
Summary.....	40
CHAPTER III: METHODOLOGY.....	42
Introduction.....	42
Participants and Distribution.....	43
Survey Questions.....	44
Survey Pilot.....	44
Research Design.....	45
Data Analysis.....	46
Summary.....	53
CHAPTER IV: RESULTS.....	54
Introduction.....	54
Descriptive Statistics.....	55
Results for Research Question 1.....	57
Results for Research Question 2.....	58
Results for Research Question 3.....	60
Results for Research Question 4.....	62
Results for Research Question 5.....	63
Conclusion.....	64

CHAPTER V: DISCUSSION.....	66
Introduction.....	66
Research Question 1.....	67
Reflection of Student Learning.....	67
Reflection of Content Taught.....	68
Data to Guide Student Learning.....	69
Research Question 2.....	70
Research Question 3.....	72
Research Question 4.....	74
Research Question 5.....	75
Conclusions.....	77
Limitations.....	79
Recommendations for Future Research and Policy Implications.....	80
REFERENCES.....	83
APPENDIX A: MAP-KCCT COMPARISON.....	95
APPENDIX B: IRB APPROVAL.....	100
APPENDIX C: INFORMED CONSENT FORM.....	101
APPENDIX D: SURVEY INSTRUMENT.....	103
APPENDIX E: CURRICULUM VITAE.....	106

LIST OF TABLES

Table 1: Kentucky Proficiency Standards in Comparison to NAEP Standards.....	5
Table 2: Lexile Score Comparison Between KCCT and MAP.....	6
Table 3: Demographic Statistics of Demographic Variables.....	56
Table 4: Comparison of Kentucky Educators’ Perceptions of KCCT Results for Mathematics and Reading.....	58
Table 5: Comparison of Kentucky Educators’ Perceptions of KCCT Results for Difficulty Level in Mathematics and Reading.....	59
Table 6: Comparison of Kentucky Educators’ Perceptions of KCCT Results for Difficulty level by Student Ability.....	59
Table 7: Comparison of Kentucky Educators’ Perceptions of KCCT Results for Appropriate Student Classification in Mathematics and Reading.....	61
Table 8: Comparison of Kentucky Educators’ Perceptions of KCCT Results for Appropriate Classification of Students.....	62
Table 9: Comparison of Kentucky Educators’ Perceptions of KCCT Results for Accuracy of Student Sub-group Academic Ability.....	62
Table 10: Comparison of Kentucky Educators’ Perceptions of KCCT Results for Accuracy of Student Sub-group Academic Ability.....	63
Table 11: Comparison of Kentucky Educators’ Perceptions of KCCT Compared to Other Instruments.....	64
Table 12: Analysis of Kentucky Educators’ Perceptions of KCCT Compared to Other Instruments.....	64

PERCEPTIONS OF KENTUCKY EDUCATORS CONCERNING THE
KENTUCKY STATE ASSESSMENT SYSTEM AS AN ACCURATE REFLECTION
OF STUDENT LEARNING

Benny C. Lile

May 2012

108 Pages

Directed by: Fred Carter, Kyong Chon, Lisa Murley, and Jerry Ralston

Educational Leadership Doctoral Program

Western Kentucky University

While educational testing has been in place since the one room school house, it was not until the 1990s that accountability began to accompany assessment programs. With the passage of the No Child Left Behind Act (NCLB) in 2001, virtually every public school district in the United States of America that desired to continue to receive Title 1 funding found themselves tied to rigorous assessment and accountability systems. This focus on accountability has impacted every school, district, and state as they have sought to implement and deal with the consequences it has wrought. As the 50 states have sought to deal with federal mandates, other countries are seeking better alternatives for national testing systems as well.

Countless data have been collected and articles written over the past decade concerning the impact and subsequent ramifications of NCLB. This study sought to bring to the discussion a missing factor, that being the voice of practitioners. Amidst the volumes of information, there is a void of hard evidence from the field.

The research sought to answer five questions: (1) What are the perceptions of Kentucky educators concerning the results of the Kentucky Core Content for Assessment (KCCT) as an accurate reflection of student learning of Mathematics and Reading in specific areas (e.g., student learning, content taught, and instructional guidance)?; (2) What are the perceptions of Kentucky educators concerning the difficulty of the KCCT

for students of different academic ability levels?; (3) What are the perceptions of Kentucky educators concerning the accuracy of student performance classification for the results of the KCCT?; (4) What are the perceptions of Kentucky educators in regard to the KCCT results being an accurate reflection in relation to student ability for the No Child Left Behind (NCLB) sub-groups of Special Education, Free/Reduced Lunch, and English as a Second Language (ESL)?; and (5) What are the perceptions of Kentucky educators concerning the accuracy of the KCCT, as opposed to other national measures?

Results indicate reservations exist concerning the accuracy of results of the Kentucky assessment system. Further study is warranted to determine the underlying causation of perceptions of Kentucky educators.

CHAPTER I: INTRODUCTION

Standardized testing has existed in America since at least the early 1920s, when Edward Thorndike and colleagues of the Columbia Teachers College developed a system to measure students' abilities in core areas, writing, and handwriting (Ravitch, 2000). As time elapsed, the education community continually tried to perfect large scale assessments to best indicate the actual skills and abilities of students. Assessments, currently being used as measures of accountability systems, began to come into vogue in the early 1990s largely as a result of the 1983 federal report, *A Nation at Risk* (United States Department of Education, 1983).

Assessment is a method of measurement of what a student knows and is able to do and provides an indication of what is not known. Accountability is defined as holding someone, or some organization(s), responsible for what has been learned. The ultimate goal of any single assessment or assessment program should be to advance student learning. Learning is evaluated in terms of an overall level rather than a single score, which serves to make the measurement consistent with the learning (Marzano, Pickering, & McTighe, 1993). Reeves (2002b) wrote that, as student learning goes much deeper than assignments and grades, the assessment that takes place within the school building also should go deeper. Evidence exists that schools have navigated the complex and sometimes demoralizing capacities of assessment and accountability systems and used them as stimulates for positive growth (Reeves, 2004). All states are accountable to the federal government for Title 1 funding based on results from state-administered

assessments. This requirement is a central tenant of PL 107-110, the reauthorization of the Elementary and Secondary Education Act, better known as No Child Left Behind (NCLB; 2001). Due to the lack of a national curriculum, individual state departments of education have been free to choose any preferred testing instruments. Some states have as many as four or more unique components within individual assessment programs (Wolff, 1998). Given the variety of assessments utilized by the different states, and the varying definitions of proficiency (considered satisfactory performance by NCLB), the clarity of student performance across the country can create an environment of confusion (Yin, Schmidt, & Besag, 2006).

The focus of this study was to determine whether the large scale state assessment model provides an accurate measure of student learning. This question has produced anxiety with staff members, bewilderment among parents, rebellion of students, and ongoing turmoil within state and national political bodies (Amrein & Berliner, 2002; Bonner, 2007; Elbousty, 2009; Perlstein, 2007). Regardless of the outcome of the study question, it is reasonable to assume that statewide assessments and some varying degree of accountability will remain in the near future (United States Department of Education, 2010).

Problem Statement

Prior to the reform efforts of the 1990s, states typically used norm-referenced testing. As statewide systems began to be developed, more criterion-based measures started to appear. Kentucky became a national leader in that regard (Steffy, 1993; Whitford & Jones, 2000). The purpose of the criterion-based model is to test against a standards-based body of knowledge. The instrument is designed to measure the mastery

level knowledge of a student based upon the standards. Norm-referenced exams are designed to compare students and groups of students across a defined norming group (Bond, 2008).

Upon entering the new millennium, a marked rise can be seen in the use of criterion-based examinations for the summative purpose of accountability, leading to more emphasis upon formative-based assessments. Quality assessment practices within the classroom have been found to provide the best, and most accurate, diagnostic feedback for a teacher (Guskey, 2003). Educators have learned that the effective use of formative instruments can lead to exemplary results on summative state assessments (Reeves, 2004).

D'Agostino, Welsh, and Corson (2007) noted that instructional practices have been affected by assessment systems in ways not necessarily considered pedagogically sound. State and federal emphasis on assessment and accountability created an emphasis on summative evaluations. Educators recently have begun to focus on formative assessments that occur throughout the year as better measures of understanding student learning (Reeves, 2004). Newton (2007) wrote that summative assessments come without purpose, and formative assessments come without judgments. Less refined teaching methods were thought to bring about better standardized test scores, while quality-rich instruction did not. This belief among staff members stemmed from the idea that a singular focus on tested content would produce a better result. Reeves (2004) posited that this is not always true.

Wagner et al. (2006) shared examples of schools heavily focused on the goal of test score improvement. The authors encouraged a much broader vision of school

improvement that included instructional practice and authentic student achievement at the core. The lack of significant energy and investments on school improvement is considered a leading weakness of today's accountability systems (Elmore, 2008).

Consideration of the effect of assessments upon students should be more important than any other aspect. The effort level on the part of the student has been one of much study and debate (Wise & DeMars, 2005). When teachers are in an environment where they feel free to provide varied instruction and assessment methods based on student need and interest, student achievement tends to flourish (Marzano, 2006). If students are presented with high quality and engaging assignments, they respond in like fashion. As stated by Reeves (2004), students can and will respond to quality teaching.

The National Center for Education Statistics (NCES) studies continually indicate that the vast majority of state measures of proficiency fail to meet the same level, as defined and assessed by the National Assessment of Educational Progress (NAEP) (NCES, 2005). Not only do these standards not match, but many states deem proficiency to be at a basic level as defined by NAEP. In addition, the definition of proficient exhibits a great variance between the different states (NCES, 2009). This data is interpreted for Kentucky in Table 1.

Table 1

Kentucky Proficiency Standards in Comparison to NAEP Standards

Subject Grade	NAEP Standard	Kentucky Standard
Mathematics – 4 th	Basic	Proficient
Mathematics – 8 th	Basic	Proficient
Reading – 4 th	Below Basic	Proficient
Reading – 8 th	Basic	Proficient

Note. Adapted from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2009 Reading and Mathematics Assessments. U.S. Department of Education, Office of Planning, Evaluation and Policy Development, EDFacts SY 2008–09, Washington, DC, 2010. The National Longitudinal School-Level State Assessment Score Database (NLSLSASD) 2010.

Additional correlation studies between Measurement of Academic Progress (MAP) and Kentucky Core Content Test (KCCT) scores are shown in Appendix A. (This information was collected from reports produced by the Northwest Evaluation Association.) In this study, Dahlin (2008) presents data indicating a student who scores in the 27th percentile in reading and the 34th percentile in mathematics has a better than average chance of scoring proficient on the Kentucky Core Content Test in each respective subject area. Correlation studies between Lexile scores produced by both MAP and KCCT are presented in Table 2. Lexile scores on both exams are produced by an approved linking study (Lexile, 2011). (This information was collected from a rural Southcentral Kentucky school district that has participated in MAP testing for over five years.)

Table 2

Lexile Score Comparison Between KCCT and MAP

Grade	KCCT AVG L Score	MAP AVG L Score	Difference	% of Students Scoring Higher On KCCT
3 rd	728	626	102	78
4 th	787	794	-7	62
5 th	937	869	68	63
6 th	1067	925	142	83
7 th	1073	1011	62	64
8 th	1108	1049	59	63

Note. Results are reported for NCLB accountable grades only. Both data sets are from spring 2011 test administrations.

The data indicates that MAP presents a more rigorous assessment for the students, as higher Lexile scores are more difficult to obtain via the MAP instrument on five out of six of the tested and reported grade levels. In all reported grade levels, the majority of individual students produced a higher Lexile score on the KCCT assessment.

Purpose and Background

While numerous statistical analyses of state assessment results have been conducted throughout the years, and a plethora of anecdotal articles exist in the media, little research has been conducted regarding practitioner perceptions concerning the results of state assessments and accountability judgments. This study seeks to identify perceptions related to the results of the Kentucky state assessment, along with additional information into the sub-components of the program. This task will be accomplished by asking nine questions in relation to the assessment results, particularly the NCLB accountable subject areas of Mathematics and Reading.

The ultimate purpose of the research will be to investigate the above mentioned factors and present the results along with the existing qualitative correlation reports. Comparing and contrasting existing data sets with practitioner perceptions will present an accurate picture of the status of the Commonwealth of Kentucky assessment and accountability system. Figure 1 indicates the varying degrees of complexity and interrelated components that constitute a thorough study of an assessment system.

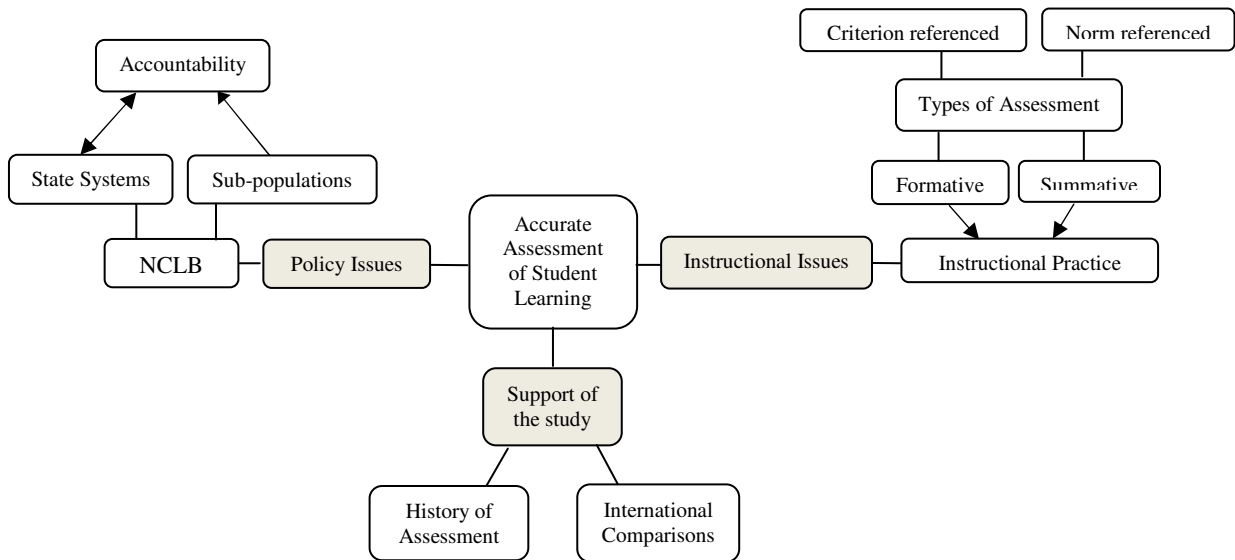


Figure 1. A graphic representation of the various components that are related to, and affect, the research of the problem.

The comparison of assessment systems and results between states has been difficult. The United States, as a nation, also provides a difficult comparison to other countries. The nation is unique to the majority of other countries because no national curriculum or national exam exists. This is being somewhat addressed by the new Common Core Standards and the consortia to build voluntary national exams (USDOE, 2010). Several aspects of assessment systems were explored in preparation for this project, including those of an international perspective. The countries of Singapore and Finland, both regarded as international leaders in education, are briefly reviewed.

Specifically, the system of assessment and accountability in Costa Rica was reviewed in depth.

International Comparisons

Finland. For the past decade, the nation of Finland has been recognized as an international leader in education. The Program for International Student Assessment (PISA) ranks nations based on those who are members of the Organization for Economic Cooperation Development (OECD) and for nations that take the exam but are not members. The latest measure of the PISA indicates Finland ranks second among the 34 OECD participating countries in Reading and Mathematics and first in Science (National Center for Education Statistics, 2011).

Finland does not have a national exam at any time during or at the end of a student's experience in the school system. All assessments are developed by classroom teachers. The teacher is held in high regard and is given a great deal of autonomy. Finland had a rigid national curriculum until the 1990s, when a series of reforms took place that brought about a great deal of flexibility. Most Finnish experts attribute the overall success of the educational system to a myriad of interrelated factors, both inside and outside of the school house (Valijarvi, Linnakyla, Kupari, Reinikainen, & Arffman, 2002).

Singapore. Much like Finland, Singapore has a reputation for scoring well on international comparison examinations. Although Singapore is not a member of the OECD, it is one of 31 other nations that participate in the PISA. In the most recent results, Singapore ranked third in Reading and Science and second in Mathematics among the non-OECD nations.

Unlike Finland, Singapore has a series of high stakes national examinations. These tests begin in the elementary school grades and will determine what path a student will take in high school. Two more series of exams during the high school years determine the type of post-secondary school a student will be able to attend (Gregory & Clarke, 2003).

Costa Rica. Costa Rica began national educational assessments in 1986 (Ferrer, 2006). Several Latin American countries began to participate in international assessment initiatives, chief among them being the Third International Mathematics and Science Survey (TIMSS). The initial results for the Latin American nations were not flattering (Wolff, 1998) and served as the impetus for many of the participating nations to place a greater emphasis on their national testing models. The TIMSS results provided a reason to explore the existence and purpose behind the national exams.

In Costa Rica, national exams (for accountability) are administered at the end of Grades 6, 9, and 12. The students are tested in Foreign Language, Language, Mathematics, Natural Sciences, and Social Sciences. The exams are considered high stakes, as they are 40% of the final grade in 6th grade and 60% for the senior. In addition, admission to the university system is based upon the results. It should be noted that participation in the national exam is required of both public and private school students (Ferrer, 2006).

The Institute for Research to Improve Costa Rican Education, which is an independent branch of the University of Cost Rica, was responsible for the initial implementation of the national assessment system (Wolff, 1998). Through the years, this responsibility changed hands several times and now lies with the Quality Control

Division of the central government's Ministry of Education (Ferrer, 2006). While a lack of consistency existed throughout the years, the current model of assessment lends itself to a more student centered approach. Diagnostic tests are found in the primary grades, with assessments for problem-solving ability and physical capacities. Even though the accountability exams are norm-referenced, the Ministry of Education produces a criterion report based on the results (Wolff, 1998).

Costa Rica has not been without problems as it sought to institute a national assessment model. In the 1980s, as they began to implement the first system of assessment across the nation, the country faced economic collapse. This affected all aspects of life, and education was not spared (Molina & Palmer, 2009). The country continued to focus on the importance of education in their advancement as a developing nation (Navarro, Carnoy, & Castro, 1999). As Costa Rica began to recover throughout the 1990s, the Ministry of Education explored new ways to measure student abilities. The use of more performance type measures was introduced but quickly abandoned due to large scale scoring errors (Wolff, 1998). Additionally, the Ministry found it difficult to define the intended audience. While reports are shared with various segments of the education community, surveys indicate there is little knowledge of the results and scant change in instructional practices. The question is left open as to whether this is a result of the political structure of the country, or if it speaks to the competence and/or concern of the country's educational officials.

Research Questions and Hypotheses

This study will determine the perceptions of Kentucky educators in regard to large scale state assessment results. For the purpose of this study perception is defined as the attitudes and beliefs of the respondents. The survey of experienced teachers and administrators will address the following research questions:

Research Question 1: What are the perceptions of Kentucky educators concerning the results of the Kentucky Core Content for Assessment (KCCT) as an accurate reflection of student learning of Mathematics and Reading in specific areas (e.g., student learning, content taught, and instructional guidance)?

Hypothesis: The average respondent will indicate the results of the KCCT for Mathematics and Reading are, at minimum, an adequate indicator of student learning. A variance of responses in terms of subject areas is expected.

Research Question 2: What are the perceptions of Kentucky educators concerning the difficulty of the KCCT for students of different academic ability levels?

Hypothesis: The average respondent will indicate the results of the KCCT for Mathematics and Reading have minimal variance among the student ability groups.

Research Question 3: What are the perceptions of Kentucky educators concerning the accuracy of student performance classification for the results of the KCCT?

Hypothesis: The average respondent will indicate little variance among the four classification groups in both the subject areas of Mathematics and Reading, and the results would be skewed toward the “Accurate” ranking.

Research Question 4: What are the perceptions of Kentucky educators in regard to the KCCT results being an accurate reflection in relation to student ability for the No

Child Left Behind (NCLB) sub-groups of Special Education, Free/Reduced Lunch, and English as a Second Language (ESL)?

Hypothesis: The average respondent will find the accuracy of the KCCT results for sub-populations comparable in nature to those of the general tested population.

Research Question 5: What are the perceptions of Kentucky educators concerning the accuracy of the KCCT, as opposed to other national measures?

Hypothesis: The average respondent will find that the KCCT results are comparable in nature to those of nationally available standardized assessments.

Significance of Study

In many regards, testing for organizational accountability is a relatively young phenomenon in the United States of America (NCLB, 2002). This study is significant, in that it seeks to explore the perceptions of those closest to the students. While an assessment instrument produces a score and/or a performance judgment on behalf of an individual student and organizational entity (i.e., school or district), the classroom teacher and other building personnel know better than anyone the capabilities of those being tested. Through the survey results, this research seeks to provide a true and clear picture of the perceptions of those practitioners.

The questions being asked go beyond a statistical comparison. They seek to bring forth a professional judgment yielded only by those who are in regular contact with those being tested. The questions probe not only the perceptions of the accuracy of the assessment results, but they seek a deeper understanding of the sub-components of the assessment process. Tested subject areas, performance judgment classifications, and student sub-population performance were all explored. Given the assessment program in

Kentucky has an annual budget exceeding ten million dollars, this study becomes even more relevant (Kentucky Office of State Budget Director, 2011). This study will allow for the professional judgments to be compared with research in order to form a more complete understanding of large scale assessment programs. Specially, it may serve to explain the gap between national test results and state test results. Ultimately, a better understanding of actual student abilities should lead researchers and practitioners to discover new and more efficient methods to increase student learning.

Definition of Terms

Authentic Learning – Schlechty (1997) discusses authenticity in terms of the opposite term, inauthentic. When students participate in activities that are perceived to be artificial and/or contrived, they are less likely to gain any lasting knowledge or understanding. Conversely, if the learning has been internalized and seems to have relevance to the student, the more likely it is to be deemed a true or real experience. In turn, the student should be able to re-create or demonstrate gained knowledge.

Common Core Standards – The Common Core Standards is an effort led by the Council of Chief State School Officers and the National Governors Association. The goal is to create a set of content standards that may be voluntarily adopted by each state in order to produce a sense of national uniformity in what students should know and be able to do. Standards for English/Language Arts and Mathematics are now complete. Presently, 45 states have adopted the standards (Common Core State Standards Initiative, 2011).

Criterion Referenced Tests – Instruments that can be described as measuring a specific body of content are known as criterion referenced tests. The object is to measure

the degree, or amount, of content that a student has learned during a specified time period. Development of these instruments is not as problematic as the process of reporting and the creation of performance judgment categories (Koretz, 2008).

Formative Assessment – Assessments that pair the efforts of the student and teacher in order to develop an individual learning progression are said to be formative in nature. Formative assessments are normally not a formal instrument and can take on many variations. The key to the formative assessment is that it charts the continued learning path for individual students (Moss & Brookhart, 2009).

High Stakes Assessment – Amrein and Berliner (2002) describe high stakes assessments as those that have consequences tied to them. In the current environment, these consequences would be synonymous with accountability. The high stakes can take many forms, including district and/or school rewards or sanctions, and grade retention for individual students.

Lexile – A Lexile is an equal interval unit of measure that is used to describe an individual student reading level. There is no official Lexile examination instrument, but an appropriate Lexile score can be taken from any number of approved tests (Lexile, 2011).

Norm Referenced Tests – Tests that are designed to provide a comparison of students one to another are considered norm referenced tests. The test publisher will use a large set of data from a specific testing time frame and will declare it the norm. Subsequent test scores will be compared against the score of this norming group in order to determine a percentile ranking for each individual student (Bond, 2008).

Standardized Tests – Koretz (2000) describes standardized tests as those instruments that provide uniformity of administration (in terms of both time and environment), questions, and scoring. All students who participate in an exam considered to be standardized would have a similar testing experience, regardless of location or individuals proctoring the test.

Summative Assessment – Assessments that provide a summary of what a student has learned are deemed to be summative in nature. They are normally not used to guide student instruction and are often used to provide a large percentage of a final grade (Guskey, 2002).

CHAPTER II: REVIEW OF THE LITERATURE

Introduction

Tests and measurement have a long history in the field of education. The field has always been rich with research and information concerning the psychometric concepts and constructs of assessment instruments. The onset of state, and now national, accountability systems has produced a plethora of data concerning not only the instruments, but related areas such as test results, instructional significance, and curriculum decisions that are influenced by the different assessment and accountability systems.

This literature review will begin by examining the basic psychometric functions of testing instruments. The review will then cover such topics as assessment results, No Child Left Behind influences, and state issues. It will conclude with a review of the effect assessment and accountability systems have on curriculum and instruction issues.

Of the varied ways to inspect assessment systems, validity and reliability are paramount in the field. The entire testing profession, from those who construct and produce instruments to the institutions that administer and interpret the results, is appropriately concerned with the concept of validity and reliability. Particularly in the area of validity, any number of tangential issues affects the interpretation of results. Nichols and Williams (2009) explored how a test score can affect an individual or institution in regard to the impact upon validity.

Nichols and Williams (2009) posited that researchers in the early 1950s began to speak of validity in three realms: content, construct, and criterion. Guion (1980) called this “...something of a holy trinity representing three different roads to psychometric salvation” (p. 386). Nichols and Williams constructed the remainder of their comments around the consequences of score use and concluded that it falls outside of Guion’s “trinity.”

Nichols and Williams (2009) cited the work of Shepard (1997) and Kane (2001) in stating, “Test developers are not responsible for the negative consequences following test score misuse or for distal consequences...” (p. 5). This is further examined in *STANDARDS for educational and psychological testing* (AERA, APA, NCME, 1999). Accepting the accuracy of this statement, the question must then be asked, “Who is responsible?” The authors contended there should be a shared responsibility of score interpretation and presented a concise graphic representation outlying this concept.

Qualitative studies using constant comparative analysis were used to measure the actions and attitudes of teachers in the state of New York regarding standardized testing. A 20-question survey was distributed to various grade level teachers in western New York (Klein, Zevenbergen, & Brown, 2006). The intent was to garner specific teacher instructional practices and attitudes concerning the state standardized tests that were given at the three grade levels: elementary, middle, and high. The team sought to address four core questions, each with underlying indicators: (1) How does testing influence teaching?; (2) How does testing affect the way students learn?; (3) How is content taught reflected in the test used?; and (4) How does testing influence views of self and education? (Klein et al., 2006).

The results were consistent with other research findings and literature reviews. While the intent of the standards-based movement and related accountability was for students to achieve mastery in relation to the stated standards, it was never meant to limit student opportunity or narrow the curriculum. Although standards have been criticized for being too broad and difficult to cover, the related assessments were too narrow and specifically focused on a small sub-set of the standards (Shmoker & Marzano, 1999; Commonwealth Educational Policy Institute, 2000). The New York research indicated a strong propensity to teach to specific test content throughout the year, while less attention was given to differentiated teaching practices and activities more suited to individual student success. The disparity was so great that nearly three quarters of those surveyed indicated they taught test content throughout the year, while less than five percent provided authentic instruction and prepared students who need extra help (Klein et al., 2006).

The research team received enough anecdotal comments on the surveys to speculate that much of the test preparation activities were done at the behest of school and district administrators. They recommended professional development and related educational initiatives for all school personnel to prove that quality instruction and high test scores can coexist.

Levitt (2008) posited in his dissertation that No Child Left Behind was inherently flawed, as well as the South Carolina state assessment system used to measure student achievement. The purpose of the study was to argue that the Measurement of Academic Progress (MAP) was a much more accurate indicator of student learning. (MAP is produced by the Northwest Evaluation Association [NWEA] and is a computerized

assessment.) He presented a quantitative study comparing scores on the South Carolina state assessment to the scores of the same students on the MAP. The intent was to formulate answers to the following four questions relative to the achievement gaps in the areas of Reading and Mathematics. (1) What do most recently available South Carolina MAP performance results from the NWEA indicate concerning the achievement gap between ethnic groups?; (2) What do the most recent MAP results indicate about the achievement gap among socio-economic sub-groups?; (3) What does the MAP performance reveal regarding the gap when considering the number of times a school had administered MAP? Finally, (4) what does the gap reveal when combining results for ethnic and socio-economic groups? Levitt concluded the MAP assessment system provided a much more accurate and precise picture of individual student achievement than the current state model and, hence, should be the instrument used to provide the judgments for the NCLB results.

Effects on Results

The research team of Wise and DeMars (2005) studied the effect of tests that bear no consequence for the student in terms of either academic credit or a grade, hence, the term low-stakes. After reviewing concepts and theories behind test-taking motivation, they tackled the effect of low motivation on an individual result and the interpretation of a group performance.

Student motivation, or test-taking effort, was described as the level of student engagement and the energy expended in order to perform as well as possible on the test (Wise & DeMars, 2005). Human nature would logically be the reason any individual would perform at a lower level when there was nothing to gain and at increasingly higher

levels when personal benefit was increased. In fact, of the 15 studies cited, only one (Kiplinger & Linn, 1992) did not produce a positive correlation between motivation and performance.

A variety of issues could influence the data beyond motivation. Students who have more knowledge very possibly will be better motivated to perform on the assessment. Thus, higher ability may lead to higher motivation. In return, students not confident of their ability are more apt to not try as hard (Wise & DeMars, 2005).

Performance issues also related to the type of test being administered. Students reported a higher motivation for multiple-choice tests that will be scored than for those not scored (Sundre, 1999). The standard deviation was found to double when presenting the same situation with an essay question (Wolf & Smith, 1995).

When trying to determine the validity of assessments that may be low stakes for the students but high stakes for an organization or institution, a conundrum is created. Generous data were provided to prove that motivation is an issue. The conclusion of the document focused on possible solutions. Wise and DeMars (2005) took a more positive approach. They explored solutions that might encourage students to give their best in terms of rewards or motivation.

In the ongoing quest to apply causative rationale to assessment results, the research team of O'Neil, Abedi, Miyoshi, and Mastergeorge (2005) explored the effect of monetary compensation on performance of a low-stakes exam. Released Mathematics items from the TIMSS (1997) were used. The released items included 12 multiple-choice questions and 8 free-response items. Students were asked to gauge their motivation on three 6-item scales (O'Neil, Sugrue, Abedi, Baker, & Golen, 1997). Adequate

preparation was made in terms of creating a control group and a motivation group, as well as allowing for variables such as gender, differing test forms, and prior academic performance. The motivation group would receive 10 dollars for each correct answer, while the control group would receive nothing. Both groups were asked to rate their level of motivation and effort at different times on the 6-item scale.

The results were quite surprising, as the motivation group performed no better than the control group (O'Neil et al., 2005). Even more surprising was that the information showed the motivation group put forth more effort but did not score better. The natural inclination would be to expect greater effort to lead to better performance, but this was not the case. The authors adequately discussed possible explanations for the findings, and were confident in the procedure, as the alpha reliability of the effort scale was .85.

If the effort and motivation were unrelated to performance, the exam may actually have been a solid measure of the degree of content knowledge retained by the student. Should that be the case, the assessment instrument performed as designed; and, although the researchers did not reach the hypothesized conclusion, it possibly brought valuable information to the testing community.

An Arizona State University research team (Amrein & Berliner, 2002) surveyed 18 states in order to evaluate the effectiveness of their respective high-stakes accountability systems relative to actual student learning. To formulate the comparison, the ACT, Scholastic Aptitude Test (SAT), National Assessment of Educational Progress (NAEP), and Advanced Placement (AP) exams were used. All were considered to have an overlapping effect with the state standards.

The research provided a brief historical review of high-stakes testing in America, which is considered to have its roots in the now defunct minimum competency movement. Amrein and Berliner (2002) are two of the few to note the impact of *A Nation at Risk* upon the high-stakes environment. A clear and commonsense outline in favor of high-stakes testing was presented. That argument is countered, however, with what Amrein and Berliner denoted as the uncertainty principle. A concept from the field of social sciences, the principle postulates that decisions become more and more corrupted and distorted in relation to the social impact that they hold. If this principle holds true, then expecting pristine results from a high-stakes environment may be nearly impossible.

In addition to the very real possibility of inaccurate results, the study speculated that schools and districts may be doing their students and community a disservice when in search of a higher score. By narrowing the curriculum and denying a wider field of study, students may be less educated when achieving a high test score number. The study ultimately pointed out that, at best, the actual level of student learning was unable to be determined. The researchers suggested continued discussion is needed at the highest policy-making levels concerning the viability of high-stakes assessments.

Sub-population Issues

Special populations can be found among the many issues that compound the area of assessment result validity. The Individuals with Disabilities Education Act (IDEA; 1997) and the NCLB Act (2002) state that all students, including those with disabilities, participate in mandatory state assessments. Given that student disabilities can range from mild speech articulation to severe/profound cognitive functioning, the challenges are

obvious in terms of evaluating the meaning of a test score for this population. The following researchers considered the perceived value of the results for these students and a method of assessment that has the promise of better and more meaningful data.

Crawford and Tindal (2006) conducted a qualitative study of teachers and principals in Oregon concerning the inclusion of students with disabilities in state assessments. The study was conducted by using a proportional stratified random sample survey with the purpose of identifying the overall knowledge of the assessment program and the usefulness of the results. A summary of the findings indicated that teachers appear to be more familiar with the policy issues surrounding special education inclusion in the Oregon assessment program than principals. Perhaps, because of their knowledge, the teachers also are more suspect of the results for the special education student. The surveys indicated that teachers to a greater extent view the results as less useful in the way they impact the instructional process than principals. Given these results, an argument can be made for continued study and potential changes in our state assessment programs of students with disabilities.

The work of Shaftel, Yang, Glasnapp, and Poggio (2005) presented the initial response to the above cited findings. The researchers studied the implementation of a modified assessment instrument used in the Kansas assessment system. A clear delineation was made between the terms accommodations and modifications. While acceptable accommodations are permitted for Kansas students, the new initiative focused on a modification of the test. The challenge arose in creating an assessment that produced valid results, indicated true achievement, and held true to federal requirements (IDEA, 1997; NCLB, 2001). In particular, fourth-, seventh-, and tenth-grade

mathematics tests were examined. A rigorous process was applied to ensure that the new assessment was of the highest quality and comprehensive in comparison to the regular assessment (Finn & Petrilli, 2000).

A sampling of the modifications included limiting steps and simplifying language, reducing the overall number of test items, and limiting the operands. The results indicated a strong internal consistency and test reliability. The majority of individual test item results showed a strong correlation to those of the regular test (Shaftel et al., 2005).

While only two states were involved in the above studies, the information could provide a strong knowledge base for the nation. The lack of confidence from those in the teaching profession and the complexity of accommodations and modifications for special populations merit further exploration of a modified assessment system (Destefano, Shriner, & Lloyd, 2001).

State Issues

Tests may produce valid and reliable results regarding the specific instrument and the content included, but a third question relative to accuracy is warranted. Are the results an accurate measure of student learning? Any number of factors may affect this question, but only limited measures exist to address such issues. One such measure has compared state assessment results to the National Assessment of Educational Progress (NAEP). NAEP has become the *de facto* measure of accuracy for all state assessments, as defined by No Child Left Behind (Iowa State Education Association, 2007). Although this comparison bears no legal or authoritative leverage, it brings comparison results into the court of public opinion, as increasing emphasis is placed on this analysis with each year's score release (NCES, 2005).

Kentucky and Maine

Lee (2007) conducted a quantitative study to compare the results on the state assessments in Mathematics from Kentucky and Maine to their respective NAEP results for 1996 and 2003. This study utilized a stratified random sample of the fourth- and eighth-grade students from each state and compared the state assessment results to the NAEP results. The states were selected because their assessment formats closely align with that of NAEP, which includes a mixture of multiple choice and constructed response questions. Each state's internal issues were taken into consideration when compiling data, such as system design changes, differing standards and cut points, and the level of consequences for the results (Lee, 2007).

With the inclusion of the above factors, the correlation study determined that both states showed a strong positive relationship between their respective state scores and their NAEP results. While this was an overall finding, sub-category findings would bear greater study. State results are normally higher than NAEP results (Education Trust, 2009). In this study, eighth-grade students in Maine scored considerably *higher* on the NAEP than on the state assessment. The Kentucky eighth-grade results were almost a mirror opposite, with the state results much higher than NAEP (Lee, 2007).

Although the overall study produced an acceptable result for both states, the sub-scores flag the need for a more in-depth review of underlying causal factors. States appear to have often accepted the results of a large scale study without considering specific nuances contained within the study itself. Lee (2007) identified at least one area of concern.

Maryland

Parke and Lane (2007) have written extensively about the Maryland State Performance Assessment (MSPA). Their study sought to discern student perceptions as to the value and impact of the state performance assessment system. The general findings were that students had an overall positive impression of the assessment structure. The results indicated that students felt classroom instruction and related activities focused on deeper reasoning as a result of the state assessment system.

In a later study, Parke and Lane (2008) sought to determine the impact of the MSPA as it related to activities in the Mathematics classroom. Specifically, the team was studying the degree of alignment that exists between the items on the assessment instrument and activities that occur on a regular basis in the classroom. Alignment seemed to occur most often in the tested grades and trended greater to instructional activities as opposed to assessment activities.

Colorado

Colorado adopted a comprehensive package of school standards and a related assessment system in the mid-1990s. For the past decade, the ACT exam has been used to measure post-secondary readiness. All 10th-grade students are required to sit for the Colorado State Assessment (CSAP), and all 11th-grade students take the ACT as part of the state accountability system. The CSAP is used as a predictor for the ACT, with applicable remediation steps taken as an intervention, if needed. Studies found that students who were deficient on the state assessment in the late elementary and early middle grades tended to be less than college ready upon high school graduation. Current

efforts are being implemented to incorporate interim and formative assessments (Lefly, Lovell, & O'Brien, 2011).

Various States

Vanfossen and McGrew (2008) reviewed the effect of NCLB on the Social Studies curriculum in a number of states. North Carolina, South Carolina, California, Texas, and Illinois all reported less instructional minutes being devoted to Social Studies topics after the implementation of NCLB. A detailed study in Indiana showed similar findings with the minutes from Social Studies being shifted to Language Arts and Mathematics, the two subjects for which schools are held accountable under federal guidelines.

The NCES (2005) periodically conducts a quantitative correlation study by mapping scores from large scale state assessments to performance on the NAEP. The NCES compares the scores used to determine proficiency and then “maps” to the NAEP proficiency scale. This is accomplished by using the percentage of students considered proficient on the respective scales and placing them in comparison. The NCES study includes the subjects of Mathematics and Reading in Grades 4 and 8 for 2007, an ongoing project since 2003 (NCES, 2009). Appropriate statistical measures are considered in accounting for measurement error and test changes. Forty-eight states were included in the study.

The study results raise questions about the accuracy of correctly identifying and classifying proficient students. The following information was revealed when comparing the respective state definitions of proficiency. No state met the level of proficiency in Grade 4 Reading, as defined by NAEP. Only one state matched the proficiency level in

Grade 8 Reading; all others fell short. Similar results were reported for Mathematics. Only one state exceeded the standard in Grade 4, and two exceeded the mark in Grade 8. Over half of the states defined Grade 4 Reading proficiency on the state exam at a level lower than the NAEP standard of basic (NCES, 2009). The depth of this study clearly delineates the variance of standards across the nation. Many have taken this as a call to support the common core standards effort (Council of Chief State School Officers, 2009). However, some disagree with this thought or with the impact of the findings of the NCES study. Andrew D. Ho of Harvard says, “If two tests don’t measure the same thing, then mapping is misleading. You can map anything onto NAEP” (Viadero, 2009, p. 16). Certainly, the depth and variety of standards pose issues; but, without question, the chasm between levels of proficient performance among the states will create questions leading to further research and study.

Functions of Test Instruments

As if a multitude of specific human variables are not enough to account for testing irregularities, an oft misunderstood, or at best under-applied, concept of regression to the mean can be added to the mix. Smith and Smith (2005) referred to it as a statistical phenomenon that is often ignored. When test makers consider the measures of error that may occur, they most often consider recent life events that students might have experienced, the maturation of students between testing events, content motivation, and the effect of the instrument and physical environment of the test setting (Smith & Smith, 2005).

Regression is a normal variation that occurs when equal variances are correlated over a normal distribution curve (Maddala, 1992). In terms of scholastic assessment, this

would mean that a student who scores at a high level in one test setting will more than likely score lower during the next administration. Conversely, a low scoring student would likely score higher at the next administration. These are respectively known as positive and negative error scores. The research team attempted to determine the measure of this effect, as opposed to what a student's true score might be. In this case, a true score is defined as the statistical expected value of an individual score (Lord & Novick, 1968).

Using the basic regression framework (Smith & Smith, 2005), the team studied scores from the California state testing program for the years 1999-2000 and 2000-2001. Smith and Smith readily assert that "...scores that have important consequences should be interpreted properly" (p. 392). The work indicated that regression to the mean was a real phenomenon for both the individual and group score. This research indicated a need for further study of the impact of regression to the mean upon high-stakes testing systems.

Koretz (2008) presented a complete and in-depth analysis of national and state testing programs. He produced multiple decades of research and analyses, and he sought to make sense and structure of our current state of assessment in America. Beginning with the history of American testing, Koretz worked his way through common misperceptions of what tests can and cannot do, misinterpretation of results, assessment definitions, and the current problems with state programs.

A methodical analysis was presented of the perception of the American public concerning testing and why, in most cases, those perceptions are wrong. An effective comparison was made between our testing programs and what the results are intended to

relay to the public and other industries, such as the airlines. In essence, when any industry or organization is determined to prove something through a statistical procedure, it usually can be done. For example, when airlines were ordered by the federal government to have more on-time flights, they merely lengthened the flight times (Koretz, 2008). When states are compelled to have more students deemed proficient on state exams, a tendency emerges to lower the standard of what is defined as proficient (Koretz, 2008). Koretz presented ample evidence to at least suggest that some, if not all, state assessment results should be treated with some degree of suspicion.

In another report, Koretz (2000) challenged the notion that state-administered standardized test results provided an accurate picture of student achievement. Particular emphasis was placed upon the implications for pass/fail status for students and for evaluative accountability for teachers. While it seems logical that the results from a well-known state-administered test would be a valid measure of student performance and a way to hold teachers accountable for their performance, it appeared nothing could be further from the truth (Koretz, 2000).

Koretz (2008) provided a brief historical background about standardized testing in America. From the earliest beginnings, he traced the ebb and flow throughout the decades and finished with our current status of high-stakes standards-based assessment programs in most all states. The core presumptions of this paper focused on the assumption of accuracy about student gains, i.e., when the public sees gains, especially to a large degree, can they be trusted? Several instances were presented that would raise doubts about the question, but the focus was a large scale study on results from Kentucky and the impact of the introduction of a new test instrument.

An in-depth study of the Kentucky results, as compared to NAEP results from 1992-1996, indicated a major instance of score inflation on the state index, while the NAEP results were largely unchanged (Koretz, 2000). While it may be argued that the two exams were measuring competing standards, the correct point is made that the Kentucky standards for assessment were drawn from national standards. More importantly, regardless of standards differences, the Kentucky score inflation was far beyond statistical significance and should have raised a flag of attention to bring about further investigation.

After three years of administering the same test, the introduction of a new test consistently yielded inferior results. An abundance of anecdotal evidence exists to support this through the years, but Koretz (2000) provided definitive proof. A number of possible explanations were offered, from teaching directly to the test, the use of outdated norms, and the actual possibility of improved student achievement. Spalding and Cummings (1998) found similar evidence related to the writing portfolio portion of the Kentucky assessment system during this time frame.

Koretz (2000) presented several recommendations in terms of addressing these issues, although he readily admitted we will never have a perfect assessment and accountability system. The proposal showing the most promise was called a hybrid model, which used the actual test results and combined them with in-depth program reviews of the school. If results and practices appeared to match, the results were considered accurate. If they did not, more independent study would be performed and consistently low performing schools would be provided the assistance they needed.

Effects on Schools and Instruction

Moller (2009) presented an international study of the existing internal tensions in the public school sector due to accountability frameworks that have been implemented. The impact of the accountability models was explored through an analysis of case studies, as conducted by the *International Successful School Principalship Project* (Day & Leithwood, 2007). The closing statement of the conclusion best set the tone for the entire document. Moeller stated, “The focus can be on raising test scores instead of serious concern about how to promote good education for all children” (p. 45).

While creating a system of educational accountability may seem like a simple task to the lay public, efforts around the globe, such as those outlined in this work, are proving it is anything but easy. The author expended a considerable effort in differentiating between accountability and responsibility and between professional and managerial accountability. Managerial accountability is in reference to one’s standing within an organization and the expectations that follow due to that standing; whereas, professional accountability pertains to an adherence of standards specific to one’s profession (Moller, 2009). This was presented in the context of a standards-based reform model. This standards model has tended to create a notion of more individual leadership or accountability where turn-around specialists are present in schools considered failing. These systems may be as problematic as the very issues they attempt to overcome (O’Day, 2002).

With accountability systems that focus primarily on one test score result, the case studies based upon the standards-based assessments indicated the results themselves become paramount to the school where the teacher’s efforts at exploration and innovation

take a back seat (Moller, 2009). The fact that schools have and continue to “overbuy” this notion of single score accountability is perplexing. Moller stated the school should not be evaluated based on marks or test scores alone because it will create a misleading picture. Elmore (2006) argued that the problem was that many schools have little knowledge about how to effectively respond to accountability policy.

The researchers Goldschmidt, Martinez, Niemi, and Baker (2007) used a quantitative, multi-level model to conduct an in-depth analysis of the multi-faceted relationship between the California High School Exit Examination and national Stanford Achievement Test, 9th Edition. The Language Arts exit exam was considered a performance event assessment, which required answers to prompts and was scored by the local school classroom teacher (Goldschmidt et al., 2007). Due to the nature of local scoring and rubric/standard interpretation, particular interest was centered on the relation of the two different assessment scores, the fairness of each exam, the predictive nature of the exams in relation to each other, and the degree to which a transfer of learning takes place.

In addition to the study, the ability of performance assessments to provide tangential formative data also was examined. This concept is somewhat related to the theory of knowledge transfer, with the supposition that students who are instructed, prepare for, and participate in performance assessments have a higher likelihood of receiving regular formative feedback as well as formulating skills and abilities that can be used on other tasks. The use of the multi-level model allowed the researchers to account for in-school variances and to compare variables down to the student level. In addition, hierarchical linear modeling was used to further account for school level variations.

The study found an acceptable level of fairness associated with the state exams, although some gender specific issues warranted greater study (Goldschmidt et al., 2007). A significant indication of transfer of knowledge between the exams was revealed, supporting the fact that students taught in a performance-based environment have the ability to perform well on a multiple-choice test. Given these findings, this research team concluded that the California performance exams were a reliable measure of student ability, particularly in Reading and Writing. Further research was recommended, especially at the classroom level, in terms of teacher affect and opportunity to learn.

The impact of state testing, specifically high-stakes accountability measures, is a reciprocal issue in the educational K-12 setting. The argument can be made that assessment methods drive instruction. By the same token, instruction can drive assessment results. State assessment systems oftentimes have been fashioned for the explicit purpose of changing instructional practices. Indeed, the statement, “Assessment drives instruction,” has become commonplace. Former Kentucky Department of Education Associate Commissioner of Assessment and Accountability Scott Trimble (personal communication, September 8, 2009) held the belief that a quality assessment program that requires students to respond in a thoughtful and constructive manner would lead to instructional practices requiring the same. If, however, teachers can utilize instructional practices that they believe lead to more favorable test results, they will be more inclined to use those practices regardless of the soundness of pedagogical value (International Reading Association, 2009). It is difficult for an assessment instrument or program to be a driving force and a monitoring force at the same time (Gong, 2009).

Instructional Practices

Vogler (2008) conducted a qualitative study of the instructional practices of high school social studies teachers in Tennessee and Mississippi. Surveys were presented to a stratified random sample of teachers in both states taking into consideration geographic location, past state assessment results, and the number of U. S. History teachers in the system. Results were presented in terms of correlation between instructional practices and teacher attitudes/beliefs about state assessment.

The study showed that Mississippi teachers tended to use instructional practices more inclined to produce results on the state test than those in Tennessee. The Tennessee teachers indicated a desire to present material in terms of what was most beneficial to student learning. This result was somewhat predictable, in that the Mississippi system of accountability was considered more high-stakes than that of Tennessee (Volger, 2008). Hence, the test results would have a more direct reflection on the teachers in Mississippi.

In addition to a difference reported in the types of instruction, the amount of time in preparation was evident as well. Over 60% of Mississippi teachers indicated using more than two months of school time to prepare for state assessments, while in Tennessee the surveys revealed only 14% of teachers used this much time for test preparation.

The study leaves several unanswered questions as to the ultimate motivation of instructional practices. It appears that the impact of high-stakes accountability has the greatest influence on teacher practice. Although both the Tennessee and Mississippi test formats were somewhat equal, and teachers responded in similar fashion as to what types of instruction were ultimately most beneficial to students, the deciding factor was the use

of state assessment results. When the stakes are high, it appears teachers often resort to the most time efficient means available for the greatest impact on raising the test score.

A number of factors should be considered when studying the validity of test score results for any instrument and/or system. One of the more difficult is the effect of instructional practices in the classroom. D'Agostino, et al. (2007) explored the instructional implications of time devoted to test content and also the style and types of instruction. The study focused on the standards-based testing system in Arizona, specifically targeting fifth-grade Mathematics.

Since the Arizona assessment is a standards-based system, a brief discussion of the national standards movement was presented. Standards-based reform models have been in vogue since the early 1990s (Jennings, 1998; Tucker & Coddling, 1998). Measuring instructional practices appears to be a simple task, with standards readily available for correlation. However, most state academic standards are vague, broad, and oftentimes interpreted by teachers in different ways (Hill, 2001).

In some studies, the authors used the terms *instructional insensitivity* and *instructional validity* interchangeably. Sensitivity is considered a core requirement for state assessments in order to make proper score inferences (D'Agostino et al., 2007). The study was designed by administering a survey to fifth-grade teachers to measure their teaching methods in relation to tested standards. Surveys were distributed evenly between classes of varying academic achievement levels and socio-economic status. The results were compared against student scores using the Rasch system analysis (Olson & Smoyer, 1993).

The emphasis by teachers of standards alone could be expected to correlate to higher test scores. However, this was not the case. The researchers found that emphasis plus alignment led to a significant positive correlation of test scores, even after adjusting for prior achievement and demographic status. While the work brought some enlightenment to the topic, it left areas for further exploration. Since this study focused on math, would the findings be the same for other academic areas? Certainly, the presentation of standards and the methods of teaching would vary from one discipline to another. The authors presented solid evidence for a limited study and piqued the interest of the reader for more in-depth research (D'Agostino et al., 2007).

Given the impact of high-stakes assessment on classroom practices, a great emphasis is placed on the issue of instructional sensitivity. The specific construction of tests items was recommended to be reviewed in terms of discrete sensitivity. A more robust emphasis on the sensitivity issue on the part of the test publishers would produce exams that bring about more well rounded and consistent results (Polikoff, 2010).

While much of the current research around large scale assessment systems focuses on the totality of results, a need is apparent for studies that fill a niche area of specific disciplines. In fact, this study begins with the presupposition that high-stakes accountability has taken the place of literacy-based assessments across the nation (Higgins, Miller, & Wegmann, 2007). As noted throughout this review, numerous studies have documented the perceived detrimental affect of high-stakes accountability systems on student learning. The research team of Higgins et al. postulated that quality writing instruction produced students with good test scores.

The work follows the implementation of proven writing strategies (6+1 method) with a group of students and anecdotally observed writing assessment results that follow. Even as the current testing environment holds great sway over many of today's instructional practices, the outcome indicated that quality results can come from a setting not dominated by test preparation activities.

The aspect of time was one of the chief indicators highlighted in the study. Literally, each moment devoted to test preparation activities is one not given to quality writing instruction or student writing production. Time is necessary, not only to allow for instruction, but even more so to allow students time to reflect, conference, and revise. The modern day test-ready environment paints a picture where students must be continually busy with something. Some may feel "think time" is wasted time. The team concluded their work by stating that a dearth of in-depth research on the topic exists, and the field is in dire need of such studies (Higgins et al., 2007).

Amrein and Berliner (2002) studied national standardized tests in 18 states. They stated, "Although many states demonstrate increases in scores on high stakes tests, transfer of learning is not a typical outcome" (p.52). The determination was made that no clear indication of student learning can be found, even when/if scores increase (Elbousty, 2009).

Several considerations were taken into account, such as state standards and the lack of national standards. Other factors, such as financial incentives for Advanced Placement exams, were presented. This leads into a discussion of the effects of student accountability; i.e., How does the score count for the student? Is it calculated in grades, GPA, transcripts?

The issue of testing methodology also is a consideration. Given the various and sundry test formats and environments across the nation, it appears to create a difficulty in standardization of comparison. It seems that some assessment systems are designed to produce a pre-determined outcome as opposed to being an accurate reflection of student learning (Koretz, 2000, 2008). Both Wolff (1998) and Yin et al. (2006) have written about the differing types of assessment instruments and systems in place across the nation. In the age of NCLB, the many variations of testing methodologies are being used to provide a universal measure of comparative data.

Reeves (2004) writes extensively of the effects of state assessments, specifically those with high-stakes accountability, on student opportunity for learning. The fact that not all aspects of student and/or school performance can be captured with a single test score is explored as well as the correlating teacher performance with test results. Reeves favors a system termed "holistic accountability," where many more factors of the school experience are included in a performance judgment. Various models were presented that include multiple measures of school characteristics to help provide a holistic overview of school performance (Reeves, 2002a).

A common theme of more current day researchers and experts in the field is the importance of student learning. That the purpose of school should be more about learning than about tests, and that authentic assessments rather than large scale standardized assessments will lead to greater learning, has begun to receive abundant press (Wiggins, 2011). Tashlik (2010) writes of a focused project in the state of New York where a collection of high schools worked on performance assessment tasks as a measure of school effectiveness. Authentic performance tasks are found to engage

students and the entire community of learners into a richer set of learning experiences more appropriate for the advanced world in which we live.

While authentic assessment has received revived publicity, formative assessment has taken on a renewed emphasis as well. Summative assessments are used to judge student performance; formative assessments are used to guide student learning (Chappuis, 2009). Teachers can expect students to perform better on all types of assessments when formative activities are used in varied and meaningful formats (Dirksen, 2011). An important aspect of the formative process is the involvement and interaction the student has with their own learning. Students and teachers become more reflective of what has been learned as opposed to what has been exposted in the name of teaching (WestEd, 2010). The team of Tauth-Nare and Buck (2011) explored formative learning and assessments in relation to problem based learning. Specifically, in the content field of science they found students to be more inquisitive of their own learning experiences and more apt to seek out and explore new avenues to increase their own opportunities. The positive impact of effective questioning, school wide culture, and student goal setting are among the outcomes of a quality implementation of formative assessments (Moss & Brookhart, 2009).

Summary

The literature concerning the development, implementation, and impact of large scale state assessments is varied and continues to grow by volumes each day. It seems evident, as NCLB has now been in effect for 10 years, that state assessments are a predominant driving force in the American classroom. This seems particularly true given the high stakes nature of the exams.

Researchers and columnists alike point to a myriad of issues that have been created by these assessment systems. It should be noted that it is not only the assessment instruments themselves that draw the critical eye of scholars, but also the accountability systems that accompany them. Taken together it is difficult to find consistent empirical evidence that these large scale assessments have brought added value to the classroom. In instances where that may be the case, other issues quickly appear to counter any positive outcomes (Reeves, 2004; Schachter, 2011).

Noted authors, such as Reeves (2002a, 2004), Marzano (2006), and Guskey (2003), speak strongly to the use of quality classroom assessments as opposed to an undue focus on summative instruments, particularly those of the large scale standardized variety. A great inconsistency can be seen in terms of the policies produced by state and federal governments and what the respected researchers and experts in the field are reporting. The most influential writers seem not to be seeking a “one must win, one must lose” solution, but truly appear to be on a mission to determine what is best for the student. The debate concerning the value of large scale assessment systems will continue into the foreseeable future. As Margaret E. Goertz of the Center for Policy Research in Education opined, “I don’t think we’ll ever have the definitive answer that high-stakes accountability, per se, is good or bad” (Viadero, 2003, p. 12).

CHAPTER III: METHODOLOGY

Introduction

The researcher developed a survey that measured the perceptions of Kentucky educators concerning the appropriateness and accuracy of state assessment results. The researcher surveyed participants for general demographic information and then asked nine questions regarding the accuracy of the Kentucky Core Content for Assessment.

The survey intends to answer the following research questions:

Research Question 1: What are the perceptions of Kentucky educators concerning the results of the Kentucky Core Content for Assessment (KCCT) as an accurate reflection of student learning of Mathematics and Reading in specific areas (e.g., student learning, content taught, and instructional guidance)?

Research Question 2: What are the perceptions of Kentucky educators concerning the difficulty of the KCCT for students of different academic ability levels?

Research Question 3: What are the perceptions of Kentucky educators concerning the accuracy of student performance classification for the results of the KCCT?

Research Question 4: What are the perceptions of Kentucky educators in regard to the KCCT results being an accurate reflection in relation to student ability for the No Child Left Behind (NCLB) sub-groups of Special Education, Free/Reduced Lunch, and English as a Second Language (ESL)?

Research Question 5: What are the perceptions of Kentucky educators concerning the accuracy of the KCCT, as opposed to other national measures?

This chapter will provide a framework for the research methodology that was used for the study. Information is presented that will detail the pool of potential survey respondents and the various methods used to distribute the survey. An explanation of the survey instrument, pilot process, and validity and reliability checks will follow. The chapter will conclude by outlining the various statistical analyses that will take place for each survey question.

Participants and Distribution

The researcher submitted materials to the Human Subjects Review Board immediately upon development of the initial survey instrument. Permission to proceed with the survey process was granted (see Appendix B). Respondents had access to all relevant informed consent statements prior to survey participation (see Appendix C).

The researcher gave Kentucky educators access to the survey by soliciting responses through the following state e-mail listservs: *All Kentucky Superintendents, All Kentucky District Assessment Coordinators, Kentucky Association for Assessment Coordinators Membership, All Kentucky Department of Education Math and English Language Arts Regional Content Specialists, All Kentucky Valley Educational Cooperative Instructional Supervisors Leadership Network, and all the Kentucky Valley Educational Cooperative English Language Arts Teacher Leader Network*. Through an agreement with the Kentucky Association for Assessment Coordinators (KAAC), the survey was available via the organization's web site.

The most recent information indicates there were approximately 51,000 certified and actively employed P-12 Kentucky educators. Of these, approximately 3,200 had one year or less experience (Legislative Research Commission, 2008). As the directions for

the survey state that only those with more than one year of teaching experience should participate, that left approximately 47,800 potential survey respondents. An initial return rate goal of 375 was set based upon a confidence rate of 95% and a confidence interval of 5% (Raosoft, 2004).

Survey Questions

A 16-question survey was designed to capture the perceptions of Kentucky educators as they relate to the research questions (see Appendix D). The first seven questions were created to gather basic demographic data for the purpose of generalizing the population. No correlation studies were planned based on this information. The survey was to be used as a description of study participants.

The subsequent nine survey items specifically addressed the research questions. Eight of the questions were built upon a four-point Likert Scale, and the remaining question was built upon a three-point Likert Scale (International Encyclopedia of Social Sciences, 2008). In each case, the question sought educator perceptions toward the accuracy, or adequacy, of the KCCT assessment results.

Each question was vetted by the Kentucky Department of Education Associate Commissioner for Assessment and Accountability, Ken Draut. The purpose was two-fold. First was the need for an accuracy check on all statements and response choices. Second, Associate Commissioner Draut's opinions were sought to detect and eliminate any biases that might exist in the questions or response options.

Survey Pilot

For the purpose of content validity, a wide variety of educators were selected to review the informed consent documents and participate in a pilot of the actual survey

instrument (Wiersma & Jurs, 2009). Pilot participants included individuals from a variety of roles. These included, but were not limited to, teachers from various grade levels and subject areas, building and district level administrators, and certified support/resource personnel from differing areas. The individual areas were selected in order to produce a pilot that would be reflective of possible respondents to the actual survey.

The survey pilot produced 92 responses. The general demographic distribution indicated no discrepancies of note. The pilot results of the nine survey questions as well produced no suspect data.

Several quality suggestions came from the pilot. Among those most often stated included issues surrounding the technical mechanics of the instrument. Others spoke to the language used in the informed consent documents, while still others helped to clarify the statements and choices in the survey itself. After all suggested changes were considered, the researcher made a judgment as to which proposed corrections would make the survey a stronger instrument. Most often the corrections consisted of a change in wording which would help to clarify the original intent of the survey question. The final survey was presented to the dissertation committee chair, methodologist, and again to KDE Associate Commissioner Ken Draut for final review. With final approval from the above parties, the instrument was prepared to go live.

Research Design

This research is designed to discern the perceptions of Kentucky educators concerning the results of the KCCT as an accurate reflection of student learning. As stated earlier, there are several components to this question, as well as a secondary

question regarding a comparison of KCCT outcomes to those of other standardized assessment instruments.

For the purpose of this study, a quantitative design was implemented. A quantitative approach was needed in order to discern the statistical analysis of the results. The qualitative method was incorporated into the research gathering due to the subjective nature of the survey questions (Wiersma & Jurs, 2009).

The expected outcomes are as follows:

- The average respondent will indicate the results of the KCCT are, at minimum, an adequate indicator of student learning and of instructional guidance. It is expected there may be a variance of responses in terms of student classification and selected subject areas.
- The average respondent will indicate KCCT provides an adequate measure of difficulty for students of varying academic abilities.
- The average respondent will indicate KCCT results are an accurate reflection in relation to student ability for the No Child Left Behind (NCLB) sub-groups of Special Education, Free/Reduced Lunch, and English as a Second Language (ESL).
- The average respondent will find that the KCCT results are comparable in nature to those of nationally available standardized assessments.

Data Analysis

The specific data analysis will vary depending upon the nature of the question. In this section, each survey question will be described in terms of the inferential statistical procedure that will be utilized. The purpose and rationale will be explained as well. It

should be noted again that the seven leading demographic questions will not be used in a correlation nature. They will serve only to describe the background, location, and experience of the survey respondents.

- Question #1 – *Do you believe the KCCT provides an accurate reflection of actual student learning?*

The question is asked for both subject areas of Mathematics and Reading.

A four-point Likert scale response is offered with response choices of Highly Inaccurate, Somewhat Inaccurate, Somewhat Accurate, and Highly Accurate. A descriptive statistical analysis will be used to interpret the significance of the Likert scale ratings. The statistical *t*-test will be used to compare the mean variance between the subject areas of Mathematics and Reading (Wiersma & Jurs, 2009). The hypothesis states there will be minimal variance between the two subject areas, and responses will be skewed toward "Accurate" response choices. The independent variable is the subject areas in question, while the dependent variable is the perception of the accurate reflection of actual student learning based upon the four-point Likert scale.

- Question #2 - *Do you believe the KCCT provides an accurate reflection of the content that has been taught in your class?*

The question is asked for both subject areas of Mathematics and Reading.

A four-point Likert scale response is offered with response choices of Highly Inaccurate, Somewhat Inaccurate, Somewhat Accurate, and Highly Accurate. A descriptive statistical analysis will be used to interpret the

significance of the Likert scale ratings. The statistical t -test will be used to compare the mean variance between the subject areas of Mathematics and Reading (Wiersma & Jurs, 2009). The hypothesis states there will be minimal variance between the two subject areas, and responses will be skewed toward "Accurate" response choices. The independent variable is the subject areas in question, while the dependent variable is the perception of the accurate reflection of the content that has been taught based upon the four-point Likert scale.

- Question #3 - *Do you believe the KCCT provides an adequate level of difficulty for different levels of students?*

The question is asked for the subject area of Mathematics in relation to the student classifications of gifted, average, and low. A three-point Likert scale response is offered with response choices of Too Easy, About Right, and Too Difficult. A descriptive statistical analysis will be used to interpret the significance of the Likert scale ratings. The statistical t -test will be used to compare the mean variance between the subject areas of Mathematics and Reading (Wiersma & Jurs, 2009). Repeated measure ANOVA will be used to determine significant findings within subject areas. The hypothesis states there will be minimal variance between the frequency of response of the three student classifications, with the choice of "About Right" being predominant. The independent variable is the student classification categories in question, while the dependent variable

is the perception of the adequate level of difficulty based upon the three-point Likert scale.

- Question #4 - *Do you believe the KCCT provides an adequate level of difficulty for different levels of students?*

The question is asked for the subject area of Reading in relation to the student classifications of gifted, average, and low. A three-point Likert scale response is offered with response choices of Too Easy, About Right, and Too Difficult. A descriptive statistical analysis will be used to interpret the significance of the Likert scale ratings. The statistical *t*-test will be used to compare the mean variance between the subject areas of Mathematics and Reading (Wiersma & Jurs, 2009). Repeated measure ANOVA will be used to determine significant findings within subject areas. The hypothesis states there will be minimal variance between the frequency of response of the three student classifications, with the choice of "About Right" being predominant. The independent variable is the student classification categories in question, while the dependent variable is the perception of the adequate level of difficulty based upon the three-point Likert scale.

- Question #5 - *Do you believe the KCCT provides an accurate classification of students into the appropriate performance categories?*

The question is asked for the subject area of Mathematics in relation to the classification of students into the performance judgment categories as the result of a student's performance on the state assessment. A four-point

Likert scale response is offered with response choices of Highly Inaccurate, Somewhat Inaccurate, Somewhat Accurate, and Highly Accurate. A descriptive statistical analysis will be used to interpret the significance of the Likert scale ratings. The statistical *t*-test will be used to compare the mean variance between the subject areas of Mathematics and Reading (Wiersma & Jurs, 2009). Repeated measure ANOVA will be used to determine significant findings within subject area. The hypothesis states there will be minimal variance between the frequency of response in regard to the four performance categories, and responses will be skewed toward the "Accurate" categories. The independent variable is the student performance judgment classification categories in question, while the dependent variable is the perception of the accuracy of classification based upon the four-point Likert scale.

- Question #6 - *Do you believe the KCCT provides an accurate classification of students into the appropriate performance categories?*

The question is asked for the subject area of Reading in relation to the classification of students into the performance judgment categories as the result of a student's performance on the state assessment. A four-point Likert scale response is offered with response choices of Highly Inaccurate, Somewhat Inaccurate, Somewhat Accurate, and Highly Accurate. A descriptive statistical analysis will be used to interpret the significance of the Likert scale ratings. Repeated measure ANOVA will be used to determine significant findings within subject areas. The

hypothesis states there will be minimal variance between the frequency of response in regard to the four performance categories, and responses will be skewed toward the "Accurate" categories. The independent variable is the student performance judgment classification categories in question, while the dependent variable is the perception of the accuracy of classification based upon the four-point Likert scale.

- Question #7 - *Do you believe the KCCT provides adequate data to guide daily instruction?*

The question is asked for both subject areas of Mathematics and Reading. A four-point Likert scale response is offered with response choices of Highly Inadequate Data, Somewhat Inadequate Data, Somewhat Adequate Data, or Highly Adequate Data. A descriptive statistical analysis will be used to interpret the significance of the Likert scale ratings. The statistical *t*-test will be used to compare the mean variance between the subject areas of Mathematics and Reading (Wiersma & Jurs, 2009). Repeated measure ANOVA will be used to determine significant findings within subject areas. The hypothesis states there will be minimal variance between the two subject areas and responses will be skewed toward "Adequate" response choices. The independent variable is the subject areas in question, while the dependent variable is the perception of the measure of adequate data to guide daily instruction based upon the four-point Likert scale.

- Question #8 - *Do you believe the KCCT provides an accurate reflection of student ability for the various NCLB defined sub-groups?*

The question is asked in relation to student groups identified as Special Education, Free/Reduced Lunch, and English as a Second Language (ESL). A four-point Likert scale response is offered with response choices of Highly Inaccurate, Somewhat Inaccurate, Somewhat Accurate, and Highly Accurate. A descriptive statistical analysis will be used to interpret the significance of the Likert scale ratings. Repeated measure ANOVA will be used to determine significant findings within subject areas. The hypothesis states there will be minimal variance between the frequency of response of the three student sub-groups and responses will be skewed toward "Accurate" response choices. The independent variable is the sub-group areas in question, while the dependent variable is the perception of the measure of accurate student ability based upon the four-point Likert scale.

- Question #9 - *In comparison to other national assessment instruments, what do you believe is the level of accuracy of the KCCT?*

Question nine is the only item that is external in nature. The question asks for a comparison to six other nationally known educational assessments. They are ACT, PLAN, EXPLORE, Iowa Test of Basic Skills (ITBS), Measurement of Academic Progress (MAP), and Think Link. A four-point Likert scale response is offered with response choices of Less Accurate, About the Same, More Accurate, and Not Applicable. The

statistical analysis will be performed by comparing the mean and standard deviation of each response. A descriptive statistical analysis will be used to interpret the significance of the Likert scale ratings. Repeated measure ANOVA will be used to determine significant findings within subject areas. The hypothesis states there will be minimal variance between the mean of responses for the six comparative assessments and will be skewed near "About the Same" and/or "More Accurate." The independent variable is external exam group, while the dependent variable is the comparison level of accuracy as reported by survey respondents.

Summary

In this chapter, the researcher presented information regarding the development and methodology of the research study. The participants, distribution method, and research design have been discussed. The various statistical methods for data analysis, along with respective hypotheses, have been presented. The research pilot, along with results, have been described and made available. Chapter 4 will follow with the results of the research survey.

CHAPTER IV: RESULTS

Introduction

This study addressed the perceptions of Kentucky educators in regard to the accuracy of results for the state educational assessment system as a reflection of student learning. The survey instrument consisted of nine questions intended to reflect the perceptions of Kentucky educators concerning differing topics associated with the results from the state assessment. All but two of the questions were separated by requesting information in the subject areas of Reading and Mathematics. These are the two content areas for which all schools and districts are held accountable under No Child Left Behind (NCLB).

The significance of this study is due to the nature of the questions and the survey respondents. Reports are available that present conflicting data in terms of state results as compared to national results (NCES, 2009). Multiple studies question the significance of state results, again based on data analysis (Elbousty, 2009; Koretz, 2008). Few, if any, studies have sought out the perceptions of those closest to the students, those being teachers and administrators. By asking direct questions of those in the field concerning the accuracy of state assessment results, this study sought to bring clarity to the plethora of data that exist on the topic.

Research Question 1: What are the perceptions of Kentucky educators concerning the results of the Kentucky Core Content for Assessment (KCCT) as an accurate

reflection of student learning of Mathematics and Reading in specific areas (e.g., student learning, content taught, and instructional guidance)?

Research Question 2: What are the perceptions of Kentucky educators concerning the difficulty of the KCCT for students of different academic ability levels?

Research Question 3: What are the perceptions of Kentucky educators concerning the accuracy of student performance classification for the results of the KCCT?

Research Question 4: What are the perceptions of Kentucky educators in regard to the KCCT results being an accurate reflection in relation to student ability for the No Child Left Behind (NCLB) sub-groups of Special Education, Free/Reduced Lunch, and English as a Second Language (ESL)?

Research Question 5: What are the perceptions of Kentucky educators concerning the accuracy of the KCCT, as opposed to other national measures?

Descriptive Statistics

The total possible population size for the study was 47,800 (Legislative Research Commission, 2008). The survey had a total return count of 390. Based on this data, the survey had a margin of error of 4.94% and a confidence level of 95%, both well within the accepted levels for educational research (Kane, 1996; Wiersma & Jurs, 2009). The complete results of the seven-item demographic portion of the survey are presented in table format for all respondents. Demographic data was collected to show a broad representation of survey respondents. It was not meant to be reflective of the survey pilot, nor was it meant to be correlated to the specific research question response items. Table 3 provides basic demographic data of the respondents. The demographic attributes

include geographic location, school setting, grade level, role representation, subject areas taught, level of administration (if appropriate), and years of experience.

Table 3

Descriptive Statistics of Demographic Variables

	<i>N</i>	Frequency (%)
Geographic area of state	390	
Northern		12 (3)
Eastern		35 (9)
Southern		137 (35)
Central		152 (39)
Western		54 (14)
School setting	389	
Urban		13 (3)
Suburban		39 (10)
Rural		337 (87)
Grade level	389	
Elementary		172 (44)
Middle		66 (17)
High		81 (21)
District		70 (18)
Role representation	389	
Teacher		255 (66)
Administrator		134 (34)
Subject (s) taught (teachers) ^a		
Language arts - reading		110 (41)
Mathematics		92 (34)
Science		47 (18)
Social Studies		24 (9)
Other		113 (42)
Administrator level	142	
Building		72 (51)
District		70 (49)
Years in education	389	
1-5		51 (13)
6-15		139 (36)
16-25		133(34)
26>		66 (17)

^aRespondents to “Subjects taught” were allowed to make more than one selection.

Results for Research Question 1

Research Question 1 asks: What are the perceptions of Kentucky educators concerning the results of the Kentucky Core Content for Assessment (KCCT) as an accurate reflection of student learning for the subject areas of Mathematics and Reading? Specific areas of emphasis include student learning, content taught, and data to guide student learning.

A four-point Likert scale was used, with the scale choice numerical equivalents represented as follows: “Highly Inaccurate/Inadequate” – 1, “Somewhat Inaccurate/Inadequate” – 2, “Somewhat Accurate/Adequate” – 3, and “Highly Accurate/Adequate” – 4. The statistical paired *t*-test analysis was performed to determine if a significant difference in the perceptions existed between the two content areas of Mathematics and Reading. A significance level of .05 was considered to find significant difference in respondents’ perceptions between the two subject areas of Mathematics and Reading.

In direct reference to research question 1, Table 4 presents the descriptive statistics and the paired *t*-test results for the perceptions of accuracy in Mathematics and Reading. Table 4 indicates there are no significant differences in the respondents’ perceptions between the subjects of Mathematics and Reading in the areas of student learning ($t = -0.44, p = 0.62$), content taught ($t = -1.22, p = 0.22$), or instructional guidance ($t = 0.65, p = 0.51$).

Table 4

Comparison of Kentucky Educators' Perceptions of KCCT Results for Mathematics and Reading

Variable	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i>
Student learning				-0.44	0.62
Mathematics	340	2.68	.69		
Reading	336	2.69	.71		
Content taught				-1.22	0.22
Mathematics	323	2.75	.71		
Reading	323	2.76	.72		
Instructional guidance				0.65	0.51
Mathematics	333	2.31	.82		
Reading	333	2.30	.82		

Results for Research Question 2

Table 5 addresses Research Question 2 in terms of assessment difficulty as it relates to different academic levels of students for both subject areas of Mathematics and Reading. The categories of gifted, average, and low were used to describe student academic ability. Respondents were asked to rate on a three-point Likert scale, with the scale choice numerical equivalents represented as follows: “Too Easy” – 1, “About Right” – 2, and “Too Difficult” – 3.

In Table 5, the paired *t*-test result indicates a significant difference ($t = 4.18$, $p = 0.00$) in the respondents' perceptions between the subjects of Mathematics and Reading. This comparison is in terms of difficulty level for students of varying academic ability.

Table 5

Comparison of Kentucky Educators' Perceptions of KCCT Results for Difficulty Level in Mathematics and Reading

Area	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i>
Difficulty				4.18	.00*
Mathematics	338	2.18	.33		
Gifted	336	1.69	.55		
Average	339	2.11	.39		
Low	339	2.72	.48		
Reading	338	2.13	.32		
Gifted	339	1.64	.50		
Average	338	2.05	.38		
Low	338	2.68	.49		

* $p < .05$

A repeated measure ANOVA procedure was used to determine if there are differences in the educators' perceptions on KCCT's difficulty level among different academic ability groups. The results of the ANOVA are presented in Table 6.

Table 6

Comparison of Kentucky Educators' Perceptions of KCCT Results for Difficulty Level by Student Ability Level

Subject	Factor	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Mathematics		179.56	1.74	103.18	513.77	.00*
	Error	116.41	579.65	.20		
Reading		185.28	1.74	106.44	572.64	.00*
	Error	108.72	584.92	.19		

* $p < .05$

The repeated measure ANOVA results presented in Table 6 indicate a significant difference ($F = 513.77, p = 0.00$) for assessment difficulty level between student

academic ability groups in the subject area of Mathematics. There is also a significant difference ($F = 572.64, p = 0.00$) for assessment difficulty level between student academic ability groups in the subject area of Reading. For both Mathematics and Reading a significant linear pattern is observed when measuring the tests of within-subjects contrast.

Results for Research Question 3

Table 7 addresses Research Question 3 in terms of accuracy of student classification by performance judgment for both subject areas of Mathematics and Reading. The Kentucky Department of Education assigned categories of novice, apprentice, proficient, and distinguished were used to describe student classification. Respondents were asked to rate on a four-point Likert scale, with the scale choice numerical equivalents represented as follows: “Highly Inaccurate” – 1, “Somewhat Inaccurate” – 2, “Somewhat Accurate” – 3, and “Highly Accurate” – 4.

In Table 7, the paired t -test result indicates no significant difference ($t = 1.94, p = 0.05$) in the respondents’ perceptions between the subjects of Mathematics and Reading. Considering the p value barely meets the minimal acceptance level of no significance, it is possible a larger sample size would indicate a significant difference. This comparison is in terms of the appropriate classification of students by performance category.

Table 7

Comparison of Kentucky Educators' Perceptions of KCCT Results for Appropriate Student Classification in Mathematics and Reading

Area	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i>
Classification				1.94	.05
Mathematics	335	2.76	.57		
Novice	337	2.73	.69		
Apprentice	333	2.73	.63		
Proficient	335	2.78	.66		
Distinguished	335	2.78	.71		
Reading	334	2.73	.58		
Novice	333	2.69	.69		
Apprentice	333	2.72	.62		
Proficient	333	2.74	.68		
Distinguished	336	2.76	.74		

A repeated measure ANOVA procedure was used to determine if there are differences in the educators' perceptions on KCCT's accuracy of performance judgment classification in both the subject areas of Mathematics and Reading. The results of the ANOVA are presented in Table 8.

The repeated measure ANOVA results presented in Table 8 indicate no significant difference ($F = 1.56, p = 0.21$) for the accuracy of performance judgment classifications in the subject area of Mathematics. There is also no significant difference ($F = 1.64, p = 0.19$) for the accuracy of performance judgment classifications in the subject area of Reading.

Table 8

Comparison of Kentucky Educators' Perceptions of KCCT Results for Appropriate Classification of Students

Subject	Factor	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Mathematics		.83	2.13	.39	1.56	.21
	Error	176.42	703.33	.25		
Reading		.94	2.07	.46	1.64	.19
	Error	188.06	675.77	.28		

Results for Research Question 4

Research Question 4 asks: What are the perceptions of Kentucky educators in regard to the KCCT results being an accurate reflection in relation to student ability for the NCLB sub-groups of Special Education, Free/Reduced Lunch, and English as a Second Language (ESL)?

Table 9 indicates the response to the accuracy of assessment results for NCLB identified sub-groups. A four-point Likert scale was used, with the scale choice numerical equivalents represented as follows: “Highly Inaccurate” – 1, “Somewhat Inaccurate” – 2, “Somewhat Accurate” – 3, and “Highly Accurate” – 4.

Table 9

Descriptive Statistics of Kentucky Educators' Perceptions of KCCT Results for Accuracy of Student Sub-group Academic Ability

Area	<i>N</i>	<i>M</i>	<i>SD</i>
Special Education	338	1.94	.80
Free-reduced lunch	338	2.44	.77
ESL	330	2.10	.80

A repeated measure ANOVA procedure was used to determine if there are differences in the educators' perceptions on KCCT's for accuracy of results in terms of

student sub-group academic ability. The results of the ANOVA are presented in Table 10.

Table 10

Comparison of Kentucky Educators' Perceptions of KCCT Results for Accuracy of Student Sub-group Academic Ability

Factor	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Student sub-group	41.18	2.0	20.59	80.79	.00*
Error	166.16	652.0	.26		

* $p < .05$

The repeated measure ANOVA results presented in Table 10 indicate a significant difference ($F = 80.79, p = 0.00$) for the accuracy of results. This comparison is in terms of student sub-group academic ability.

Results for Research Question 5

Research Question 5 asks: What are the perceptions of Kentucky educators concerning the accuracy of the Kentucky Core Content for Assessment Test as opposed to other national measures?

Table 11 indicates the response to the accuracy of the KCCT in comparison to other national testing instruments. A four-point Likert scale was used, with the scale choice numerical equivalents represented as follows: “Less Accurate” – 1, “About the Same” – 2, “More Accurate” – 3, and “No Response”. For statistical purposes, the choice of “No Response” was not assigned a numerical value.

Table 11

Descriptive Statistics of Kentucky Educators' Perceptions of the KCCT Compared to Other Test Instruments

Area	<i>N</i>	<i>M</i>	<i>SD</i>
ACT	295	1.50	.65
PLAN	269	1.56	.67
EXPLORE	275	1.63	.72
ITBS	290	1.77	.69
MAP	277	1.75	.70
Think Link	208	1.88	.58

The Analysis of Variance ANOVA findings are presented in Table 12. A significance of $<.05$ would be considered a finding of significant difference between other national testing instruments.

Table 12

Comparison of Kentucky Educators' Perceptions of the KCCT Compared to Other Test Instruments

Factor	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Instruments	10.99	3.11	3.53	12.29	.00*
Error	145.84	507.84	.29		

* $p < .05$

The repeated measure ANOVA results presented in Table 12 indicate a significant difference ($F = 12.29, p = 0.00$) when comparing the accuracy of results from the KCCT to those of other national testing instruments.

Conclusion

This study sought to present the perceptions of Kentucky educators in regard to the accuracy of the various components of the state assessment system. Chapter 4 presented data revealing the mean rankings of perceived confidence levels of the accuracy and adequacy of the KCCT. ANOVA statistical analysis also indicated

significant differences in the areas of assessment difficulty level for students of differing academic abilities, the accuracy of assessment results for NCLB identified sub-groups, and for the comparison of the KCCT accuracy to other national test instruments. Chapter 5 will further discuss these findings as well as outline the study limitations, further recommendations, and possible policy implications.

CHAPTER V: DISCUSSION

Introduction

The purpose of this study was to determine the perceptions of Kentucky educators in regard to the Kentucky state assessment system as an accurate reflection of student learning. A nine-question survey sought to determine the degree of confidence Kentucky educators have in the varying components of the assessment system and to identify areas exhibiting statistical significant differences.

The United States of America is entering into the second decade of nationwide high-stakes school accountability (NCLB, 2002). The state of Kentucky is preparing to enter into a third decade of high-stakes school accountability (Steffy, 1993). While many have sought to marginalize the systems of assessment and accountability, there is scant evidence this will occur in the near future (USDOE, 2010).

High-stakes assessment and accountability have changed schools dramatically during this time frame. It is arguable whether this change has been positive. There are reports of a pressure charged school atmosphere when in test preparation mode and then a return to normalcy upon completion of the assessments (Perlstein, 2007). Even more troubling are the recent confirmed reports of wide scale cheating scandals prevalent in specific school districts across the nation (Schachter, 2011).

The NCES (2009, 2011) data indicate that any wide scale gains on a national measure are minimal and negligible when compared internationally. This same data indicate a disparity in comparing results from state administered assessments to national

measures (NCES, 2009). It appears a given fact that state administered, and often times state authored, assessments will continue at least for the foreseeable future (USDOE, 2010).

If decisions are being made that affect districts, schools, teachers, and students, there should be the utmost confidence in the instruments and interpretation of results that are leading to these judgments. The remainder of this chapter will explore and discuss the findings based on survey results related to this issue.

Research Question 1

Research Question 1 asks: What are the perceptions of Kentucky educators concerning the results of the Kentucky Core Content for Assessment (KCCT) as an accurate reflection of student learning for the subject areas of Mathematics and Reading? Specific areas of emphasis include student learning, content taught, and data to guide student learning.

Reflection of Student Learning

The response to the assessment being an accurate reflection of student learning produced a nearly identical result for Mathematics ($M = 2.68$) and Reading ($M = 2.69$). Considering the numerical equivalent for the response “Somewhat Inaccurate” was a 2, and the numeric equivalent for the response “Somewhat Accurate” was a 3, the overall confidence of the accurate reflection of student learning is questionable.

The hypothesis was that, at minimum, the KCCT would be an adequate indicator of student learning. As a result of the survey, the prediction is rejected, as both Mathematics and Reading showed a rating below the "Somewhat Accurate" indicator.

Kentucky educators exhibited a lack of confidence in the state assessment system in terms of the results, proving an accurate reflection of student learning. In a perfect scenario, a response of “Highly Accurate” would be desired. Considering the cost and ramifications of the assessment results, “Somewhat Accurate” should be a minimal expectation. By reflecting a lack of confidence in both content areas of Mathematics and Reading, not only is the Kentucky accountability system brought into question, but the federal NCLB judgments as well.

All parties should consider the ramifications of the results of a high-stakes, large-scale state accountability system. Everything from performance judgments for individual students, sanctions for schools (which now may include removal of teachers and administrators), and broad based public perception is based upon the test results. The media report them as factual findings that go without question. Possible reasons for this lack of confidence are discussed in subsequent sections. The research findings for this section of the study point to the fact that any and all results from the state assessment system should be treated with caution.

Reflection of Content Taught

Kentucky educators responded to the question of assessment results being a reflection content taught in Mathematics ($M = 2.75$) and Reading ($M = 2.76$). Although slightly higher than the question of accuracy of student learning, the reflection of content taught also shows a confidence level below the response choice of “Somewhat Accurate.”

The hypothesis stated that the KCCT would be an accurate reflection of the content taught. Given the findings, the hypothesis is rejected, as both Mathematics and Reading reveal a rating level below the "Somewhat Accurate" indicator.

No assurance is seen that educators believe the test results are reflective of what is taking place in the classroom. This is true for both the subject area of Mathematics and Reading. Not only do Kentucky educators question the accuracy of results from the state assessments, they also doubt that whatever accuracy may exist is indicative of the classroom content that has been taught throughout the year. The Kentucky Core Content was originally meant to be a minimal sample of content that was to be taught throughout the year and would be tested in specified grades (Steffy, 1993). Apparently, the core content soon became a maximum at some schools in the tested grade levels. Even though the intent was for the assessments to be cumulative in nature (i.e., eighth-grade Mathematics was to be a compilation of what was learned in Grades 6, 7, and 8); many educators came to feel that instruction in the "off grades" wasn't valued, but only that covered during the year of the assessment. It is very possible these attitudes lead to the findings for this question.

Data to Guide Student Learning

Kentucky educators were asked to respond to the question of whether the KCCT provided adequate data in order to guide daily instruction. A response of "Highly Inadequate" was assigned a numerical value of 1, "Somewhat Inadequate" a numerical value of 2, "Somewhat Adequate" a value of 3, and "Highly Adequate" a value of 4. Descriptive statistics indicate that respondents believed both the results from the Mathematics assessment ($M = 2.31$) and the Reading assessment ($M = 2.30$) provide less than adequate information to guide daily instruction.

The hypothesis stated there would be minimal variance between the two content areas and that responses would be skewed to the "Adequate" response categories. In

terms of variance, the prediction is accepted, as the *t*-test reflected no significant differences. The prediction is rejected for the mean response rankings, as it fell below 2.5 for both Mathematics and Reading.

The goal of any assessment should be to assist teachers in guiding and developing student learning activities (Reeves, 2002a). When asked the level of guidance provided by the KCCT for informing daily instruction, Kentucky educators were not supportive in their belief that the assessment served this function. The response was closer to “Somewhat Inadequate” than it was to any other indicator.

The cumulative result of Research Question 1 indicates that Kentucky educators are suspicious of the results in terms of being an accurate reflection of student learning, content taught, and guidance of daily learning. It seems obvious that if staff members lack confidence in these three areas, there exists a real possibility that the school culture and overall learning environment stand to be adversely affected.

Research Question 2

Research Question 2 asks: What are the perceptions of Kentucky educators concerning the difficulty of the KCCT for students of different academic ability levels?

Response choices for this category were defined as “Too Easy,” with a numerical equivalent of 1; “About Right,” with a numerical equivalent of 2; and “Too Difficult,” with a numeric equivalent of 3. Student academic ability was defined by the terms gifted, average, and low. In all ability levels, the subject area of Mathematics was found to be relatively more difficult than Reading. In the content area of Mathematics (gifted $M = 1.69$; average $M = 2.11$; low $M = 2.72$), respondents believed it somewhat easy for the gifted student and significantly more difficult for the student of lower academic ability.

The content area of Reading (gifted $M = 1.64$; average $M = 2.05$; low $M = 2.68$) showed similar results.

Using the repeated measure ANOVA of tests within subject effects, there was found to be a significant difference among different academic ability groups (gifted, average, and low). In both areas, the KCCT assessment was perceived as too difficult for lower achieving students.

Based upon the findings of the difficulty level of the state assessment for students of varying abilities, the Kentucky assessment system appears to be built for the average student. Survey responses indicate the assessment is too easy for the gifted student and too difficult for the student of lower academic ability. (This is confirmed by a later research question regarding the accuracy of results for students of different sub-populations). A statistical difference was noted among the three student academic ability groupings. This should bring into question the value of information the state assessment system is providing and both ends of the academic ability spectrum.

The percentage of students who fall into each of the three categories was not addressed. Students of all ability levels have specific needs. Gifted and talented and special education students would be considered chief among all groups. Special attention should be paid to these findings because of the unique student needs. These results possibly occurred due to the fact the KCCT is not a leveled test. On a leveled test, a student begins usually just below grade level and progresses in difficulty until a proximal point of performance is reached. Even though Kentucky reports scores in performance judgment categories from a low of novice to a high of distinguished, all students have participated in an equally difficult grade level exam. If leveled exams were used it would

truly test the upper limits of even the gifted student and would reflect the accurate ability of the special education student. Current plans are for the next generation national assessments to be a leveled exam. This is a positive step that may alleviate a portion of these concerns.

Research Question 3

Research Question 3 asks: What are the perceptions of Kentucky educators concerning the accuracy of student performance classification for the results of the KCCT?

When the state of Kentucky planned the original assessment and accountability system, the Kentucky Department of Education developed a four-tiered description of student performance (Steffy, 1993). Students are placed into the respective classification based on the KCCT results for each tested content area. A survey question was developed to solicit Kentucky educators' perceptions concerning the accuracy of this classification system. Scale ranking were "Highly Inaccurate" – 1, "Somewhat Inaccurate" – 2, "Somewhat Accurate" – 3, and "Highly Accurate" – 4.

In both subject areas, the accuracy of classification was considered better as students reached higher academic standings. As the results were quite similar for all classifications in both subject areas, the statistical *t*-test did not indicate a significant difference ($t = 1.94, p = 0.05$). To determine if there would be little variance among the four classification groups, the repeated measure ANOVA was conducted. The ANOVA results suggest that a difference was not found, however significant differences might occur with a larger sample size.

The hypothesis stated there would be little variance among the four classification groups and the results would be skewed toward the “Accurate” ranking. The mean for both subject areas was below an average of "Somewhat Adequate." For this reason, the prediction for accuracy of classification is rejected.

In all classification categories for both the content areas of Mathematics and Reading, respondents indicated a mean ranking of less than “Somewhat Accurate.” Again, responses fail to reach what should be considered a minimal level.

As stated earlier, many components of the Kentucky testing system have been in place since the inception of the program. While the performance categories have been tweaked throughout the years, they have been a constant of the program. An early criticism of the categories was the breadth of the lower two and the difficulty to achieve the highest level. This was addressed in 1998-99 with the new Commonwealth Accountability and Testing System (CATS) assessment program. The categories of novice and apprentice were divided to include the sub-categories of high, medium, and low. While this served as an acknowledgement of the issue, it still did not affect the large number of students who fell within the categories. Neither did it address the variance of academic ability that existed among students who fell at the extreme of these two categories. Score cut points were adjusted and made it easier for students to score at the highest level, which is distinguished. Even with these changes, it seems evident there are still questions concerning the appropriate classification of students into these performance judgment categories.

Research Question 4

Research Question 4 asks: What are the perceptions of Kentucky educators in regard to the KCCT results being an accurate reflection in relation to student ability for the No Child Left Behind (NCLB) sub-groups of Special Education, Free/Reduced Lunch, and English as a Second Language (ESL)?

The federal NCLB (2002) defined numerous sub-groups of students who had to meet the same achievement targets on state assessments as those of the general population at large. Although there are a number of different sub-populations defined, this study focused on the three most prevalent in Kentucky.

Survey respondents were asked to rate the accuracy of assessment results for the defined sub-populations. Scale ranking were “Highly Inaccurate” – 1, “Somewhat Inaccurate” – 2, “Somewhat Accurate” – 3, and “Highly Accurate” – 4. The results for special education ($M = 1.94$), free/reduced lunch ($M = 2.44$), and ESL ($M = 2.10$) all indicated a less than accurate reflection of student ability. The ANOVA reflected a significant difference ($F = 80.79, p = 0.00$) among the rankings for the sub-populations.

The hypothesis stated there would be minimal variance between the groups, and the scale responses would be skewed toward an accurate reflection of student ability. Due to the fact the ANOVA indicated a significant difference between groups and all mean survey responses were below 2.5, the prediction is rejected in both cases.

Statistical differences were noted between the identified sub-groups. All three of the sub-groups fell closer to “Somewhat Inaccurate” reflection of student ability. For the category of special education, the responses actually fell below the “Somewhat Inaccurate” rating. Research Question 4 provides a second confirmation that there is a

severe lack of confidence in what the state assessment results are indicating about students who fall outside what is considered an average student.

Sub-group performance may be argued to be at the heart of NCLB. At the very least, it was partially responsible for the enactment of the law. As with all of the previously addressed research questions, the magnitude of the appropriate classification of students in these sub-groups is paramount. It should be noted that a myriad of factors exert pressures upon the special education sub-population. There are any number of federal laws and regulations emanating from the department of education, civil rights, and disabilities to name a few, that can hold influence on what takes place concerning the education of these children. The possibility exists that a school may have over 20 sub-groups; and, if any one of these groups fails to meet adequate yearly progress, the school may be considered for sanctions. Considering these serious ramifications, the utmost confidence in the sub-group classification is vital.

Research Question 5

Research Question 5 asks: What are the perceptions of Kentucky educators concerning the accuracy of the KCCT as opposed to other national measures?

The state of Kentucky has gone through numerous changes of the assessment system since its inception in 1992. During this time, the core instrument of the assessment system has remained a customized, state created document. Any number of national test instruments are available. This survey question sought to compare the level of accuracy of the KCCT against other national testing instruments.

Scale rankings were classified as “Less Accurate” – 1, “About the Same” – 2, “More Accurate” – 3, and “No Response.” For statistical purposes, the choice of “No

Response” was not assigned a numerical value. The KCCT was indicated as being less accurate than all other measures. (ACT $M = 1.5$; PLAN $M = 1.56$; EXPLORE $M = 1.63$; ITBS $M = 1.77$; MAP $M = 1.75$; Think Link $M = 1.88$). The repeated measure test of within subject effects ANOVA reflected a significant difference ($F = 12.29, p = 0.00$) among the rankings for the different test instruments.

The hypothesis stated there would be minimal variance between the different test instruments, and the scale responses would be skewed toward “About the Same” and “More Accurate.” Due to the fact the ANOVA indicates a significant difference between groups and all of the six survey responses were lower than “About the Same,” the prediction is rejected in both cases.

A significant difference was noted between the six national test instruments listed. It is apparent that Kentucky educators have more confidence in all the listed national instruments than in the KCCT. Only one instrument, the ITBS, was closer to the scale ranking of “About the Same” than to “Less Accurate.” The other five instruments were closer to “Less Accurate.”

It is possible the ITBS ranked closer to the KCCT in terms of accuracy of results due to the fact it has been part of the testing system for the past two years. Even though the exam was given separately and the scores reported independently of the KCCT, it was still considered to be a part of the Kentucky assessment and accountability system.

Even though the previous research questions produced results that should cause policy makers great concern, perhaps Question 5 more than any other indicates the possible inadequacy of the program. The cost of the Kentucky assessment system is well established. Many school districts pay from local funds to administer the exams (in

addition to the KCCT) listed in this research question. The logical question follows: Why should districts have to pay to administer what they consider to be a *better* exam in terms of accurate results? The argument may be made that in an effort to produce and provide a customized assessment instrument for the state, policy makers have allowed limitations such as cost controls to override the benefits of a locally administered exam.

Conclusions

Kentucky, as have all states, has made a significant investment of time, effort, and money in the state assessment system. Upon reflection to the inception of the system in 1992, even though there have been multiple alterations, the state has probably stayed as true as any state to the original intent of the assessment system (Steffy, 1993). During the past 20 years, clarion calls have occurred for change to different components of the system. Most notable of these was the deletion of the writing portfolio (for accountability purposes) in 2009. Problems with the writing portfolio system were highlighted in years prior to this (Spalding & Cummings, 1998).

Throughout the previous two decades, most of the commentary on the state assessment system has come from those outside of the classroom walls. While teachers and administrators were free to comment and serve on task forces, committees, etc. (and many did), more often than not, the voices of concern or dissension were relegated to anecdotal incidences. Little, if any, evidence can be found of a comprehensive study of the state assessment system that took into account the perceptions of Kentucky teachers and administrators. This study has sought to do that.

Two conclusions of note should come from this work. First, this study explored the results of the assessment instrument as perceived by educators, not the instrument or

the system itself. It is up to the end user for the final determination of the ultimate quality of the results and how they are to be used. This is aligned with the definition of validity as stated in *STANDARDS for educational and psychological testing* (AERA, APA, NCME, 1999). What has led to and/or created results that lack educator confidence will be explored in a later section. Second, the results of this study should not be taken as a repudiation of the state assessment system. Legitimate issues have been raised that bear further exploration.

The Kentucky General Assembly continues to spend millions of dollars each year on the state assessment program (Kentucky Office of State Budget Director, 2011). This study indicates that Kentucky educators have reservations about the legitimacy and the value of the results. This is evidenced by the fact that not one category out of nine survey questions reaches the level of “Somewhat Accurate or Adequate.” Even though the KCCT meets the technical and operational term of valid and reliable, the concept of accuracy of results as perceived by practitioners has been brought into question. This can be further elaborated in terms of the reliability of results as a basis of interpretation by the user (AERA, APA, NCME, 1999).

The survey results from this study are consistent with the national literature base. Specifically, Wagner et al. (2006) and Koretz (2000) point to schools that focus to the point of obsession on test scores, but in reality have little actual student achievement to show for it. When considered along with national and international test comparisons, it appears there are issues concerning the results of the current system. Kentucky is not alone in this dilemma. Chapter 2 of this study highlighted more than one other state that is grappling with many similar issues. Not only are states dealing with the assessment and

accountability conundrum, but nations are as well. Even those nations that are considered high achieving are struggling to produce new and better systems.

Limitations

Existing limitations were found that may have had an effect on this study. The survey was made available to every educator throughout the state via e-mail list serve announcements and web sites. Even though it was made available on an equitable basis, the demographic data indicate a greater representation from those who identify themselves as being from the southern or central part of the state. It is possible these respondents were more familiar with the researcher and were more predisposed to complete the survey. Even with this taken into consideration, there is reason to believe the results can be generalized to the population at large.

The survey was administered electronically via the internet. Safeguards were in place that would not allow the same computer to participate in the survey more than one time. This does not, however, prevent an individual from accessing another computer(s) and participating in the survey multiple times.

Even though the survey underwent a pilot trial period, it is possible some respondents may have misunderstood or misinterpreted a question. This could in turn lead to an unintended response.

The informed consent and informational documents asked that only certified Kentucky teachers and administrators with more than one year of experience participate in the survey. An expectation of the honor system is the only safeguard to keep unqualified individuals from participating.

Recommendations for Future Research and Policy Implications

The research questions focused on the accuracy and adequacy of the results of the Kentucky state assessment system. It did not seek to find underlying causation for the results. A future study could include educator interviews around the research questions that would probe for a deeper understanding of the results. These interviews could not only probe the current research questions, but could expand to other areas as well. Some of the questions to be addressed in the future might include the following: Are inappropriate activities (test preparation, accommodations, cheating) producing results that are suspect? Has accountability produced a singular focus on a narrowed content? To what degree are other content areas suffering due to the current accountability system? This study focused on the NCLB reported content areas of Mathematics and Reading. A reasonable assumption can be made that there is a wealth of beliefs and opinions concerning other subject areas as well.

An extension of this research could assess the impact these findings have on the classroom. This study indicated educators do not have confidence in the state assessment results as an accurate reflection of student learning, the ability of the results to inform daily instruction, the classification of students, the difficulty level for varying student abilities, or as a reflection of content taught. Moreover, they believe there are any number of instruments that can do a better job of providing this information. With all of this being the case, it is reasonable to assume an atmosphere of less than ideal conditions exists in the classroom. Teacher and school morale must be in question.

An external research question is in order as to why the results from national instruments are considered superior to the Kentucky exam. Large national databases

exist that can provide a wealth of tangible information. This, coupled with educator interviews, would be a worthy study. For a more detailed examination of the assessment, analysis of the student inputs and/or their achievement data need to be included in future studies.

The research opportunities are numerous. As research continues in the field, teacher and administrator perceptions and opinions should be taken into consideration.

Just as there are any number of potential research opportunities, so too are there policy implications. As earlier stated, Kentucky spends vast sums of money each year to support a state assessment system. The strength of this study is in the fact that it presents the direct perceptions of practitioners. It appears the state is supporting a system in which its own educators lack confidence.

Continued high-stakes testing and accountability is an almost certainty. All states must continue with assessment systems at least in the immediate future as per federal mandate. Every Kentucky citizen, and especially Kentucky educators, should be concerned with the value and quality of the results being produced by the current assessment system and testing instruments.

This research suggests an underlying lack of confidence in the results of the past system. Similar to the work of Klein et al. (2006), the impact on the overall instructional process is brought into question as well. An in-depth qualitative analysis could serve to shed light on the root causes of the results that were revealed in this study. As next generation assessments and continued accountability loom on the horizon, it will be beneficial to develop a comprehensive understanding of this study and continue further

research in order to develop and maintain a system that will produce more trustworthy results.

REFERENCES

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, and NCME) (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Amrein, A., & Berliner, D. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, 10(18), 1-70.
- Bond, L. (2008). *Norm-and criterion-referenced testing*. Retrieved from <http://www.nagc.org/index.aspx?id=314>
- Bonner, C. (2007). *From coercive to spiritual: What style of leadership is prevalent in k-12 public schools?* (Doctoral dissertation). Retrieved from <http://idea.library.drexel.edu/handle/1860/1516>
- Chappuis, J. (2009). *Seven strategies of assessment for learning*. Boston, MA: Pearson.
- Common Core State Standards Initiative. (2011). *Mission statement and state adoption*. Retrieved from <http://www.corestandards.org/>
- Commonwealth Educational Policy Institute. (2000). *High stakes testing*. Retrieved from http://www.cepi.vcu.edu/policy_issues/saa/high_stakes.html
- Council of Chief State School Officers. (2009). *Common core state standards initiative*. Washington, DC. Retrieved from http://www.ccsso.org/federal_programs/13286.cfm

- Crawford, L., & Tindal, G. (2006). Policy and practice – Knowledge and beliefs of education professionals related to the inclusion of students with disabilities in a state assessment. *Remedial and Special Education, 27*(4), 208-217.
- D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*(1), 1-22.
- Dahlin, M. (2008). *A study of the alignment of the NWEA RIT scale with the Kentucky Commonwealth Accountability Testing System*. Northwest Evaluation Association, Lake Oswego, OR. Retrieved from <http://www.nwea.org/sites/www.nwea.org/files/reports/KY%20SP07%20Alignment%20 Study.pdf>
- Day, C., & Leithwood, K. (Eds.). (2007). *Successful principal leadership in times of change. An international perspective*. New York, NY: Springer.
- Destefano, L., Shriner, J. G., & Lloyd, C.A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessment. *Exceptional Children, 68*, 7-22.
- Dirksen, D. (2011). Hitting the reset button: Using formative assessment to guide instruction. *Phi Delta Kappan, 92*(7), 26-31.
- Education Trust. (2009). *Education Watch 2009 State Summary Reports*. Washington, DC. Retrieved from <http://www2.edtrust.org/edtrust/summaries2009/states.html>
- Elbousty, Y. (2009). In E. Stone, P. Swerdzewski, & M. Ewing (Co-Chairs). *High stakes literature review and critique*. Paper presented at Northeastern Educational Research Association (NERA) Annual Conference, Rocky Hill, CT.

- Elmore, R. (2006). *Leadership as the practice of improvement*. Paper presented at the International Conference on Perspectives on Leadership for Systemic Improvement. Sponsored by the Organization for Economic Cooperation and Development (OECD). London, England.
- Elmore, R. (2008). Leadership as the practice of improvement. In B. Pont, D. Nusche, & D. Hopkins (Eds.), *Improving school leadership volume 2: Case studies on system leadership*, pp. 37-67. Retrieved from <http://www.oecd.org/dataoecd/6/50/41686550.pdf>
- Ferrer, G. (2006). *Educational assessment systems in Latin America: Current practice and future challenges*. Washington, DC: PREAL.
- Finn, C. E., & Petrilli, M. J. (Eds.) (2000). *The state of state standards 2000*. Thomas B. Fordham Foundation. Retrieved from <http://www.edexcellence.net/publications/soss2000.html>
- Goldschmidt, P., Martinez, J., Niemi, D., & Baker, E. (2007). Relationships among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment*, 12(3), 239-266.
- Gong, B. (2009, November). Symposium conducted at Race to the Top, assessment program public and expert input meetings. Atlanta, GA.
- Gregory, K., & Clarke, M. (2003). High stakes assessment in England and Singapore. *Theory into Practice*, 42(1), 66-74.
- Guion, R. M. (1980). On trinitarian conceptions of validity. *Professional Psychology*, 11, 385-398.

- Guskey, T. (2002). *How's my kid doing?: A parent's guide to grades, marks, and report cards*. San Francisco, CA: Jossey-Bass.
- Guskey, T. (2003). Using data to improve student achievement. *Educational Leadership*, 60(5), 6-11.
- Higgins, B., Miller, M., & Wegmann, S. (2007). Teaching to the test...not! Balancing best practice and testing requirements in writing. *The Reading Teacher*, 60(4), 310-319.
- Hill, H. C. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal*, 38, 298-318.
- Individuals with Disabilities Education Act Amendment of 1997, Pub. L. No. 105-17, 37 Stat. 111 (1997).
- International Encyclopedia of Social Sciences. (2008). Retrieved on October, 24, 2011, from <http://www.encyclopedia.com/doc/1G2-3045301356.html>
- International Reading Association. (2009). "Template writing" found on high stakes test in Florida. Newark, DE. Retrieved from http://www.reading.org/General/Publications/blog/BlogSinglePost/09-07-21/Template_writing_found_on_high-stakes_tests_in_Florida.aspx
- Iowa State Education Association. (2007). *NAEP and NCLB testing: Confirming state test results*. Des Moines, IA. Retrieved from <http://www.isea.org/hot/accountability/naep-accountability.html>
- Jennings, J. F. (1998). *Why national standards and tests? Politics and the quest for better schools*. Thousand Oaks, CA: Sage.

- Kane, M. (1996). The precision of measurements. *Applied Measurement in Education*, (9)4, 355-379.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342.
- Kentucky Office of State Budget Director. (2011). *2010-12 Biennial Budget Revised by the 2011 Special Session*. Retrieved from <http://www.osbd.ky.gov/NR/rdonlyres/076B53F4-5359-46C0-A1BC-446B33DEA243/0/1012BOCVolumeIA.pdf>
- Kiplinger, V. L., & Linn, R. L. (1992, April). *Raising the stakes of test administration: The impact on student performance on NAEP*. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles. (ERIC Document Reproduction Service No. ED378221)
- Klein, A., Zevenbergen, A., & Brown, N. (2006). Managing standardized testing in today's schools. *Journal of Educational Thought*, 40(2), 145-157.
- Koretz, D. (2000). Limitations in the use of achievement tests as measures of educator's productivity. *The Journal of Human Resources*, (37)4, 752-777.
- Koretz, D. (2008). *Measuring up*. Cambridge, MA: Harvard University Press.
- Lee, J. (2007). Do national and state assessments converge for educational accountability? A meta-analytic synthesis of multiple measures in Maine and Kentucky. *Applied Measurement in Education*, 20(2), 171-203.

- Lefly, D., Lovell, C., & O'Brien, J. (2011). *Shining a light on college remediation in Colorado: The predictive utility of the ACT for Colorado and the Colorado student assessment program (CSAP)*. Retrieved from <http://www.cde.state.co.us/cdegen/downloads/Shiningalightonremediation2-28-2011.pdf>
- Legislative Research Commission. (2008). *Research Report No. 360. School district data profiles school year 20-07-2008*. Frankfort, KY. Retrieved from <http://www.lrc.ky.gov/lrcpubs/RR360.pdf>
- Levitt, E. J. (2008). *An analysis of student academic growth: The use of Measures of Academic Progress in South Carolina* (Doctoral Dissertation, University of South Carolina). Retrieved from <http://proquest.umi.com.libsrv.wku.edu/pqdweb?index=0&did=1574531461&SrcHMode=2&sid=2&Fmt=6&VInst=PROD&VType=PQD&RQT=309&VName=PQD&TS=1329342456&clientId=1449>
- Lexile. (2011). *Can I compare Lexile reader measures from more than one reading test?* Retrieved from <http://www.lexile.com/faq/>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores*. Reading, MA: Addison-Wesley.
- Maddala, G. S. (1992). *Introduction to econometrics* (2nd ed.). New York: Macmillan.
- Marzano, R. (2006). *Classroom assessment and grading that works*. Alexandria, VA: ASCD.
- Marzano, R., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes*. Alexandria, VA: ASCD.

- Molina, I., & Palmer, S. (2009). *The history of Costa Rica*. San Jose, Costa Rica: UCR.
- Moller, J. (2009). School leadership in an age of accountability: Tensions between managerial and professional accountability. *Journal of Education Change, 10*, 37-46.
- Moss, C., & Brookhart, S. (2009). *Advancing formative assessment in every classroom: A guide for instructional leaders*. Alexandria, VA: ASCD.
- National Center for Education Statistics. (2005). *Important Aspects of No Child Left Behind Relevant to NAEP*. Washington, DC. Retrieved from <http://nces.ed.gov/nationsreportcard/nclb.asp>
- National Center for Education Statistics. (2009). *Mapping state proficiency standards onto NAEP scales: 2005-2007*. Retrieved from <http://nces.ed.gov/nationsreportcard/pdf/studies/2010456.pdf>
- National Center for Education Statistics. (2011). *Highlights From PISA 2009*. Washington, DC. Retrieved from <http://nces.ed.gov/pubs2011/2011004.pdf>
- Navarro, J., Carnoy, M., & Castro, C. (1999). *Education reform in Latin America*. MacMillan Center for International and Area Studies. New Haven, CT. Retrieved from http://74.125.155.132/scholar?q=cache:WMfiXBiqsM8J:scholar.google.com/&hl=en&as_sdt=0,18
- Newton, P. (2007). Clarifying the purposes of educational assessment. *Assessment in Education, 14*(2), 149–170.

- Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice*, 3-9.
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).
- O'Day, J. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, 72(3), 293-330.
- Olson, A., & Smoyer, S. (1993). Local achievement testing. *Rasch Measurement Transactions*. Retrieved from <http://www.rasch.org/rmt/rmt64i.htm>
- O'Neil, H.F., Jr., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment*, 10(3), 185-208.
- O'Neil, H., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golen, S. (1997). *Final report of experimental studies on motivation and NAEP test performance* (CSE Tech. Rep. No. 427). Los Angeles: University of California, Center for Research on Evaluation, Standards and Testing.
- Parke, C., & Lane, S. (2007). Students' perceptions of a Maryland state performance assessment. *The Elementary School Journal*, 107(3), 305-324.
- Parke, C., & Lane, S. (2008). Examining alignment between state performance assessment and mathematics classroom activities. *The Journal of Educational Research*, 101(3), 132-146.
- Perlstein, L. (2007). *Tested: One American high school struggles to make the grade*. New York, NY: Holt.
- Polikoff, M. (2010). Instructional sensitivity as a psychometric property of assessment. *Educational Measurement: Issues and Practice*, 29(4), 3-14.

- Raosoft (2004). *Sample size calculator*. Retrieved from <http://www.raosoft.com/samplesize.html>
- Ravitch, D. (2000). *Left back*. New York, NY: Simon & Schuster.
- Reeves, D. (2002a). *Holistic accountability: Serving students, school, and community*. Thousand Oaks, CA: Corwin Press.
- Reeves, D. (2002b). *The daily disciplines of leadership*. San Francisco, CA: Jossey-Bass.
- Reeves, D. (2004). *Accountability for learning*. Alexandria, VA: ASCD.
- Schachter, R. (2011). Taking the helm in cheating scandals. *District Administration*, (47)10, 50-54.
- Schlechty, P. (1997). *Inventing better schools: An action plan for educational reform*. San Francisco, CA: Jossey-Bass.
- Shaftel, J., Yang, X., Glasnapp, D., & Poggio, J. (2005). Improving assessment validity for students with disabilities in large-scale assessment programs. *Educational Assessment*, 10(4), 357-375.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-8, 13, 24.
- Shmoker, M., & Marzano, R. (1999). Realizing the promise of standards-based education. *Education Leadership*, 56(6), 17-21.
- Smith, G., & Smith, J. (2005). Regression to the mean in average test scores. *Educational Assessment*, (10)4, 377-399.
- Spalding, E., & Cummings, G. (1998). It was the best of times. It was a waste of time: University of Kentucky students' views of writing under KERA. *Assessing Writing*, (5)2, 167-199.

- Steffy, B. (1993). *The Kentucky education reform. Lessons for America*. Lancaster, PA: Technomic.
- Sundre, D. L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED432588)
- Tashlik, P. (2010). Changing the national conversation on assessment. *Phi Delta Kappan*, 91(6), 55-59.
- Tauth-Nare, A., & Buck, G. (2011). Assessment for learning. *Science Teacher*, 78(1), 34-39.
- Third International Mathematics and Sciences Study. (1997, September). *Performance assessment in IEA's Third International Mathematics and Science Study (TIMSS)*. International Association for the Evaluation of Educational Achievement. Chestnut Hill, MA: Boston College.
- Tucker, M. S., & Coddling, J. B. (1998). *Standards for our schools: How to set them, measure them, and reach them*. San Francisco, CA: Jossey-Bass.
- United States Department of Education. (1983). *A nation at risk*. Retrieved from <http://www2.ed.gov/pubs/NatAtRisk/risk.html>
- United States Department of Education. (2010). *U. S. Secretary of Education Duncan announces winners of competition to improve student assessments*. Retrieved from <http://www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse>

- Valijarvi, J., Linnakyla, P., Kupari, P., Reinikainen, P., & Arffman, I. (2002). *The Finnish success in PISA - and some reasons behind it: PISA 2000*. Retrieved from <http://eric.ed.gov/PDFS/ED478054.pdf>
- Vanfossen, P., & McGrew, C. (2008). Is the sky really falling?: An update on the status of social studies in the K-5 curriculum in Indiana. *International Journal of Social Education, 23*(1), 139-179.
- Viadero, D. (2003). Researchers debate impact of tests. *Education Week, 22*(21), 1, 12.
- Viadero, D. (2009). Test rigor drops off study finds. *Education Week, 29*(10), 1, 16.
- Vogler, K. E. (2008). Comparing the impact of accountability examinations on Mississippi and Tennessee social studies teachers' instructional practices. *Educational Assessment, 13*, 1-32.
- Wagner, T., Kegan, R., Lahey, L., Lemons, R., Garnier, J., Helsing, D.,...Rasmussen, H. (2006). *Change leadership: A practical guide to transforming our schools*. San Francisco, CA: Jossey-Bass.
- WestEd. (2010). Formative assessment: Not just another test. *R&D Alert, Education Digest, 11*(2).
- Whitford, B., & Jones, K. (2000). *Accountability, assessment, and teacher commitment – Lessons from Kentucky's reform efforts*. Albany, NY: SUNY.
- Wiersma, W., & Jurs, S. (2009). *Research methods in education – An introduction, 9th ed.* Boston, MA: Pearson.
- Wiggins, G. (2011). Moving to modern assessments. *Phi Delta Kappan, 92*(7), 63.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

- Wolf, L. F., & Smith, J.K. (1995). The consequences of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, 8, 227-242.
- Wolff, L. (1998). *Educational assessments in Latin America: Current progress and future challenges*. PREAL No. 11.
- Yin, R., Schmidt, R., & Besag, F. (2006). Aggregating student achievement trends across states with different tests: Using standardized slopes as effect sizes. *Peabody Journal of Education*, 81(2), 47-61.

APPENDIX A: NWEA RIT SCALE– KCCT ALIGNMENT

A Study of the Alignment of the NWEA RIT Scale with the Kentucky Commonwealth Accountability Testing System

Michael P. Dahlin, Ph.D.

May, 2008



Copyright © 2008 Northwest Evaluation Association

All rights reserved. No part of this document may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission from NWEA.



Northwest Evaluation Association
5885 SW Meadows Road, Suite 200
Lake Oswego, OR 97035-3526

www.nwea.org
Tel 503-624-1951
Fax 503-639-7873

NWEA Kentucky State Standards Alignment Study provided courtesy of Northwest Evaluation Association. © 2012 Northwest Evaluation Association. All rights reserved.

Table 1 – Minimum Estimated Same-Season (Spring) RIT Cut Scores Corresponding to CATS Performance Levels – Mathematics

Grade	Novice			Apprentice			Proficient		Distinguished	
	Cut score	Cut score	Percentile	Cut score	Percentile	Cut score	Percentile	Cut score	Percentile	
2	<175	175	10	187	38	197	71			
3	<184	184	9	198	36	207	68			
4	<195	195	12	205	34	216	66			
5	<201	201	12	214	38	228	75			
6	<204	204	12	220	40	234	74			
7	<207	207	12	225	41	239	72			
8	<212	212	13	230	40	246	75			

Note: the cut scores shown in this table are the minimum estimated scores. Meeting the minimum MAP cut score corresponds to a 50% probability of achieving that performance level. Use the probabilities in Tables 7-12 to determine the appropriate "target" scores for a desired level of certainty.

Note: **bolded, italicized text denotes extrapolated cut score**

Table 2 – Minimum Estimated Same-Season (Spring) RIT Cut Scores Corresponding to CATS Performance Levels – Reading

Grade	Novice			Apprentice			Proficient		Distinguished	
	Cut score	Cut score	Percentile	Cut score	Percentile	Cut score	Percentile	Cut score	Percentile	
2	<156	156	2	179	25	201	82			
3	<164	164	2	189	24	210	81			
4	<175	175	4	198	27	217	82			
5	<180	180	3	204	27	222	81			
6	<180	180	2	208	27	226	79			
7	<185	185	3	212	30	230	80			
8	<183	183	2	215	28	234	82			
10	<191	191	2	223	37	241	89			

Note: the cut scores shown in this table are the minimum estimated scores. Meeting the minimum MAP cut score corresponds to a 50% probability of achieving that performance level. Use the probabilities in Tables 7-12 to determine the appropriate "target" scores for a desired level of certainty.

Note: **bolded, italicized text denotes extrapolated cut score**

Table 7 –Estimated Probability of scoring as Proficient or Higher on the CATS Mathematics Test in Same Season (Spring), by Student Grade and RIT Score Range on MAP Mathematics

RIT Range	Estimated Probability of Passing State Test for Student with Given RIT Score						
	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
125	0%	0%	0%	0%	0%	0%	0%
130	0%	0%	0%	0%	0%	0%	0%
135	1%	0%	0%	0%	0%	0%	0%
140	1%	0%	0%	0%	0%	0%	0%
145	2%	1%	0%	0%	0%	0%	0%
150	3%	1%	1%	0%	0%	0%	0%
155	5%	2%	1%	0%	0%	0%	0%
160	8%	3%	1%	1%	0%	0%	0%
165	12%	4%	2%	1%	1%	0%	0%
170	18%	7%	4%	1%	1%	1%	0%
175	27%	11%	6%	2%	1%	1%	1%
180	38%	17%	9%	4%	2%	1%	1%
185	50%	25%	14%	6%	4%	2%	1%
190	62%	36%	22%	10%	6%	4%	2%
195	73%	48%	31%	16%	9%	6%	4%
200	82%	60%	43%	23%	14%	9%	6%
205	88%	71%	55%	33%	22%	14%	9%
210	92%	80%	67%	45%	31%	22%	14%
215	95%	87%	77%	57%	43%	31%	22%
220	97%	92%	84%	69%	55%	43%	31%
225	98%	95%	90%	78%	67%	55%	43%
230	99%	97%	94%	86%	77%	67%	55%
235	99%	98%	96%	91%	84%	77%	67%
240	100%	99%	98%	94%	90%	84%	77%
245	100%	99%	99%	96%	94%	90%	84%
250	100%	100%	99%	98%	96%	94%	90%
255	100%	100%	99%	99%	98%	96%	94%
260	100%	100%	100%	99%	99%	98%	96%
265	100%	100%	100%	99%	99%	99%	98%
270	100%	100%	100%	100%	99%	99%	99%
275	100%	100%	100%	100%	100%	99%	99%
280	100%	100%	100%	100%	100%	100%	99%
285	100%	100%	100%	100%	100%	100%	100%
290	100%	100%	100%	100%	100%	100%	100%
295	100%	100%	100%	100%	100%	100%	100%
300	100%	100%	100%	100%	100%	100%	100%

Note: This table provides the estimated probability of passing the state test based on a MAP test score taken during that same (spring) season. Example: if a third grade student scored 170 on a MAP test taken during the spring season, her/his estimated probability of passing the state test is 7%.

Table 10 – Estimated Probability of scoring as Proficient or Higher on the CATS Mathematics Test Based on Prior Season (Fall) MAP Score, by Student Grade and RIT Score Range on MAP Mathematics

Estimated Probability of Passing State Test for Student with Given RIT Score							
RIT Range	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
125	1%	0%	0%	0%	0%	0%	0%
130	1%	0%	0%	0%	0%	0%	0%
135	2%	1%	0%	0%	0%	0%	0%
140	4%	1%	0%	0%	0%	0%	0%
145	6%	1%	1%	0%	0%	0%	0%
150	9%	2%	1%	0%	0%	0%	0%
155	14%	4%	1%	1%	0%	0%	0%
160	22%	6%	2%	1%	1%	0%	0%
165	31%	10%	4%	2%	1%	0%	0%
170	43%	16%	6%	3%	1%	1%	0%
175	55%	23%	10%	4%	2%	1%	1%
180	67%	33%	16%	7%	4%	2%	1%
185	77%	45%	23%	11%	6%	3%	2%
190	84%	57%	33%	17%	9%	5%	3%
195	90%	69%	45%	25%	14%	8%	5%
200	94%	78%	57%	36%	22%	13%	8%
205	96%	86%	69%	48%	31%	20%	13%
210	98%	91%	78%	60%	43%	29%	20%
215	99%	94%	86%	71%	55%	40%	29%
220	99%	96%	91%	80%	67%	52%	40%
225	99%	98%	94%	87%	77%	64%	52%
230	100%	99%	96%	92%	84%	75%	64%
235	100%	99%	98%	95%	90%	83%	75%
240	100%	99%	99%	97%	94%	89%	83%
245	100%	100%	99%	98%	96%	93%	89%
250	100%	100%	99%	99%	98%	96%	93%
255	100%	100%	100%	99%	99%	97%	96%
260	100%	100%	100%	100%	99%	98%	97%
265	100%	100%	100%	100%	99%	99%	98%
270	100%	100%	100%	100%	100%	99%	99%
275	100%	100%	100%	100%	100%	100%	99%
280	100%	100%	100%	100%	100%	100%	100%
285	100%	100%	100%	100%	100%	100%	100%
290	100%	100%	100%	100%	100%	100%	100%
295	100%	100%	100%	100%	100%	100%	100%
300	100%	100%	100%	100%	100%	100%	100%

Note: This table provides the estimated probability of passing the state test in spring, based on a MAP test score taken during the previous (fall) season. Example: if a third grade student scored 170 on a MAP test taken during the fall season, her/his estimated probability of passing the state test in spring is 16%.

APPENDIX B: INSTITUTIONAL REVIEW BOARD APPROVAL



A LEADING AMERICAN UNIVERSITY WITH INTERNATIONAL REACH
HUMAN SUBJECTS REVIEW BOARD

In future correspondence, please refer to HS12-009, July 26, 2011

Benny Lile
c/o Dr. Fred Carter
Educational Leadership
WKU

Benny Lile:

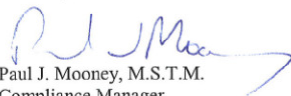
Your research project, *Teacher's Perceptions of Kentucky State Assessment as an Accurate Reflection of Student Learning*, was reviewed by the IRB and it has been determined that risks to subjects are: (1) minimized and reasonable; and that (2) research procedures are consistent with a sound research design and do not expose the subjects to unnecessary risk. Reviewers determined that: (1) benefits to subjects are considered along with the importance of the topic and that outcomes are reasonable; (2) selection of subjects is equitable; and (3) the purposes of the research and the research setting is amenable to subjects' welfare and producing desired outcomes; that indications of coercion or prejudice are absent, and that participation is clearly voluntary.

1. In addition, the IRB found that you need to orient participants as follows: (1) signed informed consent is not required; (2) Provision is made for collecting, using and storing data in a manner that protects the safety and privacy of the subjects and the confidentiality of the data. (3) Appropriate safeguards are included to protect the rights and welfare of the subjects.

This project is therefore approved at the Exempt from Full Board Review Level.

2. Please note that the institution is not responsible for any actions regarding this protocol before approval. If you expand the project at a later date to use other instruments please re-apply. Copies of your request for human subjects review, your application, and this approval, are maintained in the Office of Sponsored Programs at the above address. Please report any changes to this approved protocol to this office. A Continuing Review protocol will be sent to you in the future to determine the status of the project. Also, please use the stamped approval forms to assure participants of compliance with The Office of Human Research Protections regulations.

Sincerely,


Paul J. Mooney, M.S.T.M.
Compliance Manager
Office of Research
Western Kentucky University



cc: HS file number Lile HS12-009

The Spirit Makes the Master

Office of Sponsored Programs | Western Kentucky University | 1906 College Heights Blvd. #11026 | Bowling Green, KY 42101-1026
phone: 270.745.4652 | fax: 270.745.4211 | email: paul.mooney@wku.edu | web: <http://www.wku.edu/Dept/Support/SponsPrg/grants/index.php?page=research-compliance>
Equal Education and Employment Opportunities • Printing paid from state funds, RDS 57.375 • Hearing Impaired Only: 270.745.5369

APPENDIX C: INFORMED CONSENT

ASSESSMENT RESEARCH

The data from this survey is being collected in order to satisfy dissertation research requirements for the Ed.D degree program for Benny Lile. The topic of the research is, **"The perceptions of Kentucky educators concerning the results of the Kentucky Core Content for Assessment Test (KCCT) as an accurate reflection of student learning."**

There are no known risks for your participation in this research study. There is no signed informed consent statement available or necessary for this study. The information collected may not benefit you directly, but will contribute to other research of this topic and will better inform the profession on the issue of state assessments. The information you provide will assist in providing clarity as to the usefulness of large scale assessments.

Your response will be completely confidential. The survey contains no personal information and participation in the survey is completely anonymous. Taking part in this study is voluntary. **If you are not a certified Kentucky educator with at least one year of teaching or administrative experience, please do not participate.** By completing this survey you agree to take part in this research study. You do not have to answer any questions which make you uncomfortable. You may choose not to take part at all. If you decide to be in this study you may stop taking part at any time. If you decide not to be in this study or if you stop taking part at any time, you will not lose any benefits for which you may qualify. If you have any questions or concerns about the research study, please contact: Benny Lile at 270-651-3787 or Dr. Fred Carter (committee chairperson) at 270-745-4897. Additionally, you may call the WKU Compliance Manager at (270) 745-2129, regarding your rights as a research participant.

Refusal to participate in this study will have no effect on any future services you may be entitled to from Western Kentucky University. Anyone who agrees to participate in this study is free to withdraw from the study at any time with no penalty.

All data and information collected is used strictly for the purposes of research and analysis for the benefit of this dissertation project. Data collection and storage will protect the safety and privacy of all participating subjects as well as the confidentiality of the data. All appropriate safeguards are included to protect the rights and welfare of the subjects.

By clicking on the link below you are indicating you understand and agree with this informed consent document.

SURVEY INSTRUCTIONS

To survey respondent,

Please reflect upon each survey question in terms of general applicability to your whole class or appropriate group of students. **The initials KCCT always stand for the “Kentucky Core Content Test”.**

This survey is completely anonymous and cannot be tracked to any individual. You are asked to complete the demographic information at the beginning so the results can be better generalized in terms of geographic, grade level, and subjects taught.

When interpreting the response rubric please consider the following example definitions;

Highly – Should apply when the description is occurring in the vast majority of instances.

Somewhat- Should apply when the description may occur on a reasonably regular basis.

The survey consists of 16 total questions. There are 7 demographic questions and 9 research questions. It should take no more than 10 minutes to complete.

[Please click here in order to participate in the survey.](#)

APPENDIX D: SURVEY INSTRUMENT

qualtrics.com

1. What geographical area of the state do you consider your school district?

Northern

Eastern

Southern

Central

Western

2. Do you consider your school district to be?

Urban

Suburban

Rural

3. Please indicate your school level.

Elementary

Middle

High

District

4. Are you a(n)....

Teacher

Administrator

5. If you indicated teacher, please identify the subject(s) you teach (if multiple subjects please list the two you teach the majority of the time.)

Language Arts/Reading

Math

Science

Social Studies

Other

6. If you indicated administrator, are you....

Building Level

District Level

7. How many years have you been in education?

1-5

6-15

16-25

26 or longer

>>

Survey Powered By [Qualtrics](#)

Do you believe the results of the KCCT provide an accurate reflection of actual student learning?

	Highly Inaccurate	Somewhat Inaccurate	Somewhat Accurate	Highly Accurate
Mathematics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reading	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you believe the KCCT provides an accurate reflection of the content that has been taught in your class?

	Highly Inaccurate	Somewhat Inaccurate	Somewhat Accurate	Highly Accurate
Mathematics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reading	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you believe the KCCT provides an adequate level of difficulty for different levels of students in the subject area of Mathematics?

	Too Easy	About Right	Too Difficult
Gifted	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Average	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Low	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you believe the KCCT provides an adequate level of difficulty for different levels of students in the subject area of Reading?

	Too Easy	About Right	Too Difficult
Gifted	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Average	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Low	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you believe the KCCT provides an accurate classification of students into the appropriate performance categories for the subject area of Mathematics?

	Highly Inaccurate	Somewhat Inaccurate	Somewhat Accurate	Highly Accurate
Novice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Apprentice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Proficient	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Distinguished	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you believe the KCCT provides an accurate classification of students into the appropriate performance categories for the subject area of Reading?

	Highly Inaccurate	Somewhat Inaccurate	Somewhat Accurate	Highly Accurate
Novice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Apprentice	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Proficient	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Distinguished	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you believe the results of the KCCT provide adequate data to guide daily instruction?

	Highly Inadequate Data	Somewhat Inadequate Data	Somewhat Adequate Data	Highly Adequate Data
Mathematics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reading	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



Do you believe the results of the KCCT provide an accurate reflection of student ability for the various NCLB defined sub-groups?

	Highly Inaccurate Data	Somewhat Inaccurate Data	Somewhat Accurate Data	Highly Accurate Data
Special Education	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Free/Reduced Lunch	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
English as Second Language (ESL)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

In comparison to other national assessment instruments, what do you believe is the level of accuracy for the KCCT?

	Less Accurate	About The Same	More Accurate	Not Applicable
ACT	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
PLAN	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
EXPLORE	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Iowa Test of Basic Skills (ITBS)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Measurement of Academic Progress (MAP)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Think Link	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

You may go back and review your answers at this time by clicking the "<<" button below. When you click the ">>" button below, your survey responses will be submitted and recorded. Thank you for completing this survey.



<< >>

APPENDIX E: CURRICULUM VITAE

Benny C. Lile

Personal Data:

Name: Benny C. Lile
Address: 1955 Roberts Road, Hardyville, Kentucky 42746
Telephone: (270) 565-1762 home, (270) 651-3787 office
E-mail: lile@scrtc.com

Education:

- Western Kentucky University** – Bowling Green, Kentucky
- Doctor of Education – May 2012
- Western Kentucky University** – Bowling Green, Kentucky
- Superintendent Certification – 1997
 - Master of Arts – Education – 1988
 - Middle Grades Certification, Social Studies and Science – 1985, 1987
- University of Kentucky** – Lexington, KY
- Bachelor of Science – Agriculture Education – 1983

Experience:

- 1994 – Present – Barren County Schools** – Glasgow, KY
Director of Instruction and Technology
- 1992 – 1994 – Kentucky Department of Education** – Frankfort, KY
Regional Technology Coordinator (30 districts)
- 1985 – 1992 – Metcalfe County Schools** – Edmonton, KY
Middle Grades Teacher – North Metcalfe Elementary – Science,
Social Studies, Computer

Recognition and Membership:

- **Kentucky Association for Assessment Coordinators** – Two terms as President
- **Kentucky Association of Technology Coordinators** – Kentucky Technology Leader of the Year, President
- **Kentucky Department of Education** – Member and subsequent chair of the Governor appointed *School Curriculum, Assessment, and Accountability Council*

- **United States Department of Education** – Race to the Top assessment symposium, public and expert input meetings – Participant
- **Council of Chief State School Officers** – National high school summit; Annual policy forum – Panel participant
- **Education Commission of the States** – State Forum on Educational Accountability – Participant
- **Kentucky Department of Education** – Member of the *Kentucky Virtual Leadership* advisory council
- **Kentucky Association of School Administrators** – Regional Representative Board Member
- **International Society of Technology Educators** – NETS Technology Standards for Teachers Writing Team
- **ASCD** – Member
- **MENSA** – Member
- **Student Technology Leadership Program** – Outstanding Ambassador
- **Google Certified Administrator**
- **Kentucky Department of Education** – Member of the state middle school task force for performance-based assessment
- **Kentucky PTA** – First place award for “Effective methods of teaching the United States Constitution”
- **North Metcalfe Elementary School** – Member of the inaugural School Based Decision Making Council

Presentations:

- **Kentucky Next Generation Innovation Summit** – *BAVEL* – *Building a public virtual high school* – December 2011
- **International Center for Leadership in Education, Model Schools Conference** – *BAVEL* – *Not all schools have walls* – June 2011
- **International Center for Leadership in Education, Model Schools Conference** – *Achieve 3000* – *Accelerating reading achievement* – June 2007
- **Microsoft Connected Learning Community Technology Summit** – *Technology resource teachers, your portal to student success* – February 2002
- **National School Boards Association** – *Technology Salute District* – November 2001
- **ASCD Annual Conference** – *State assessment panel discussion* – March 2001
- **National Education Computing Conference** – *Building a district-wide technology resource program* – June 1998
- **National School Board Association Technology + Learning Conference** – *Technology in the middle school* – November 1995

- **Kentucky Department of Education-** *State-wide technology coordinator workshops, “Technology and multiple intelligence”* – 1993
- **Kentucky School Boards State Convention -** *Technology in the rural school* – February 1991
- **Kentucky Educational Technology Conference** – *Effective computer software for the at-risk student* – March 1990

Publications:

- Lile, B. (2009, January). It’s more than bits and bytes! *South Central Kentucky Business Journal*. 2009.
- Lile, B. (2004, January). Anecdotes not yardstick for school testing – guest column. *Lexington Herald-Leader*.
- Lile, B. (1990). Effective at-risk interventions, *The Link*. Appalachian Educational Resources.
- Lile, B. (1989, August). Kentucky schools at a crossroads – guest editorial. *The Courier-Journal*.

