Criterion-Referenced Assessment Literacy of Educators

James David King
University of Southern Mississippi

Follow this and additional works at: https://aquila.usm.edu/dissertations
Part of the Educational Leadership Commons, Elementary and Middle and Secondary Education Administration Commons, and the Other Teacher Education and Professional Development Commons

Recommended Citation
King, James David, "Criterion-Referenced Assessment Literacy of Educators" (2010). Dissertations. 980.
https://aquila.usm.edu/dissertations/980

This Dissertation is brought to you for free and open access by The Aquila Digital Community. It has been accepted for inclusion in Dissertations by an authorized administrator of The Aquila Digital Community. For more information, please contact aquilastaff@usm.edu.
CRITERION-REFERENCED ASSESSMENT LITERACY OF EDUCATORS

by

James David King

Abstract of a Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

August 2010
ABSTRACT

CRITERION-REFERENCED ASSESSMENT LITERACY OF EDUCATORS

by James David King

August 2010

This study investigated the assessment literacy of educators in regard to criterion-referenced tests. For the purpose of this study, administrators included principals and assistant principals within a school. First, the study determined the degree of training in measurement, assessment, and statistics that educators have received. Second, the study investigated the assessment knowledge of educators based on their performance on the Criterion-Referenced Questionnaire. Third, the study investigated if years experience, type of certification, or highest degree held had a significant difference on their performance on the Criterion-Referenced Questionnaire. Fourth, the study investigated educators' attitudes toward the use of measurement, assessment, and statistics in education.

In addition, this study sought to investigate assessment literacy of educators and help to identify areas that educators need for professional development to become more competent in the use of assessment. Educators could use the instruments from this study to help identify their own needs for professional development. This research could also help to determine future needs in assessment training for teachers and administrators in college and other certification programs.
Three separate one way ANOVAs were used for hypotheses testing in this research project. The ANOVA for years experience was not significant, $F(4, 375) = 2.41, p = .049$. Eta Square was .025. The Bonferroni Correction to control for Type 1 error was .016; thus, there was no significant difference between the groups. The ANOVA for certification was not significant, $F(1, 378) = 3.649, p = .057$. Eta Square was .009. The ANOVA for highest degree held was significant, $F(2, 377) = 11.275, p < .001$. Eta square was .056. A significant ANOVA was followed up by a post hoc test using a Bonferroni Correction to control for Type 1 error $p < .001$. 
The University of Southern Mississippi

CRITERION-REFERENCED ASSESSMENT LITERACY OF EDUCATORS

by

James David King

A Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved:

Dr. Thelma J. Roberson

________________________________________
Director

Dr. Rose M. McNeese

________________________________________

Dr. David E. Lee

________________________________________

Dr. Kyna J. Shelley

________________________________________

Dr. James T. Johnson

________________________________________

Dr. Susan A. Siltanen

________________________________________

Dean of the Graduate School

August 2010
DEDICATION

The writer of this dissertation would like to dedicate this paper to my father-in-law and mother-in-law, James D. and Elizabeth M. Smith. I would not have completed this process without their encouragement and support for which I will always be grateful.
ACKNOWLEDGMENTS

The writer would like to thank the dissertation chairperson, Dr. Thelma Roberson, and the other committee members, Dr. Kyna Shelley, Dr. James T. Johnson, Dr. David Lee, and Dr. Rose McNeese for their help on this research project. I would especially like to thank Dr. Thelma Roberson for help in the editing process and solving many problems that occurred during the completion of this dissertation. I would like to thank Dr. Kyna Shelley and Dr. James T. Johnson for helping to develop my interest in educational statistics.
TABLE OF CONTENTS

ABSTRACT.................................................................................................................... ii

DEDICATION................................................................................................................ iv

ACKNOWLEDGMENTS................................................................................................. v

LIST OF TABLES.......................................................................................................... viii

CHAPTER

I. INTRODUCTION................................................................. 1

Background
Statement of the Problem
Research Questions
Purpose of Study
Hypotheses
Definitions
Delimitations
Assumptions
Justification

II. REVIEW OF RELATED LITERATURE................................................. 18

History of Large Scale Assessments
Types of Large Scale Assessments
Theoretical Framework of Large Scale Assessment
Assessment Literacy of Teachers and Administrators
Importance of Assessment Literacy
Misuses of Assessment
Summary

III. METHODOLOGY.............................................................................. 59

Summary of Study
Research Questions
Participants
Instrument
Procedure
Analysis
IV. ANALYSIS OF DATA.................................................................69
   Instrument
   Results
   Hypotheses Testing Results

V. SUMMARY................................................................................96
   Conclusions
   Limitations
   Recommendations for Policy
   Recommendations for Future Research

APPENDIXES.................................................................................114

REFERENCES..............................................................................129
LIST OF TABLES

Table
1. Demographic Information of Participants..........................................................70
2. Descriptive Statistics Likert Items........................................................................73
3. Pattern Matrix Criterion-Referenced Questionnaire..............................................76
4. Descriptive Statistics for Job Type........................................................................78
5. Descriptive Statistic Criterion-Referenced Questionnaire........................................79
6. Descriptive Statistics Measurement Courses.........................................................81
7. Descriptive Educational Statistics Courses...........................................................82
8. Descriptive Math Statistics Courses by Job............................................................83
9. Descriptive Statistics Research Design Courses.....................................................84
10. Descriptive Statistics Measurement Courses Certification....................................85
11. Descriptive Educational Statistics Courses Certification........................................86
12. Descriptive Math Statistics Courses Certification..................................................87
13. Descriptive Statistics Research Design Courses Certification..............................88
14. Descriptive Statistics Measurement Courses Degree............................................89
15. Descriptive Educational Statistics Courses Degree...............................................90
16. Descriptive Math Statistics Courses Degree........................................................91
17. Descriptive Statistics Research Design Courses Degree........................................92
18. Descriptive Statistic Degree/Years Experience.....................................................95
CHAPTER I
INTRODUCTION

Background

The accountability movement has greatly influenced how schools test their students. According to Hursh (2005), the modern accountability movement started in the United States in 1983 with the publication of *A Nation at Risk*. Since that time the federal government’s role in education has increased in importance. With the enactment of the No Child Left Behind Act (NCLB), P.I. 107-110, 20 U.S.C. 6301 (2001), standardized testing has increased in importance. The law mandated certain testing requirements for all public school districts in the United States. Under NCLB students must be tested annually in grades 3 through 8 and at least once in grades 10th through 12th to determine if they have met proficiency in reading and math (Hursh). An assessment in science was required by the 2007-2008 school year for grades 3 through 8 (Daugherty, 2004). Furthermore, the law mandates that all students be proficient in each area by the year 2014 (Hanson, Burton, & Guam, 2006).

State Departments of Education and school districts have developed tests to meet the requirements of NCLB and according to Daugherty (2004), two main types of tests coming into general use are: comprehensive tests and end-of-course exams. Comprehensive tests assess mastery of objectives in required courses previously taken by the student (Daugherty). End-of-course exams measure a student’s mastery of the curriculum for particular courses and are given immediately following the completion of a particular course (Daugherty). In 2004, Daugherty reported that 9 southeastern states, including Alabama, use
comprehensive exams in order to comply with NCLB. Examples of such tests include: The Alabama Reading and Mathematics Test (ARMT), a criterion-referenced comprehensive exam, and The Alabama High School Graduation Exam, also a criterion-referenced comprehensive exam, designed to measure competency before a high school diploma is awarded. Daugherty asserted that end-of-course exams are better at getting teachers to be more consistent in providing instruction aligned with state standards. Ten states currently administer end-of-course-exams, according to the United States Department of Education (2009).

Increased accountability has led to development of new educational standards. Many states require state and national tests to ensure that students master basic skills (Nichols, 2003). As of 2005, twenty-four states required a high school exit examination to receive a diploma (Perking-Gough, 2005). Twenty states required students to pass a comprehensive exit exam to receive a diploma in 2004, compared to 19 states in 2003 (Gayler & Kober, 2004). Five additional states will implement an exit exam by 2009 (Chudowsky, Gayler, Hamilton, & Kober, 2002). According to the United States Department of Education (2009), twenty-five states require high school exit examinations in order to receive a diploma as of the 2007-2008 school year.

General characteristics of exit exams are: calculator use (math), time limits, release of practice items and answers, and score reporting and feedback (Gayler & Kober, 2004). English and math are the most commonly tested subject areas (Gayler & Kober). According to the Gayler and Kober, only one-third of the
states required sub area testing in science and social studies in 2002. Clearly, the number of states testing science will increase because of the mandate in No Child Left Behind. Gayler and Kober, stated only 10 states tested science in 2004 compared to 7 in 2002, and 9 states tested social studies compared 5 in the same period of time. The only state to test computer skills was North Carolina (Gayler & Kober).

High stakes testing is expensive and may include up to 80% in hidden costs (Gayler & Kober, 2004) associated with prevention, remediation, and professional development. In a Center of Education Policy study, Gayler and Kober calculated costs for implementing exit exams in high schools and reported a wide range of costs from state to state. For instance, in 2004, Minnesota spent $171.00 per pupil for its graduation exam; Massachusetts spent $385.00 per pupil for its graduation exam; and Indiana spent $557.00 per pupil on its Graduation Qualifying Exam. Their study also found that local systems pay 96% of the current cost of these new tests. The difference was about $280.00 per pupil. The expenses paid by local systems take money away from other programs. The direct cost of testing such as developing, administering, and scoring the test is a small percent of the cost. According to Gayler and Kober, less than one-fifth of the cost of the exit exam is used to administer and score the test.

The Alabama High School Graduation Exam (AHSGE) grew out of the Education Accountability Law of 1995, passed by the state legislature (ASDE, 2004). The Alabama State Department of Education determined there was a
need for higher standards in order to make Alabama students more competitive in the workplace so in 1996 the Alabama State Department of Education passed a new “4 x 4 curriculum” in which all students must complete four years in required core areas such as mathematics, science, social studies, and English. Alabama students, since 2004, must earn 24 credits of coursework including the 4 x 4 curriculum and pass the Alabama High School Graduation Exam to receive a high school diploma (ASDE). The Alabama High School Graduation Exam (AHSGE) is a standard based exam (ASDE, 2003). The Alabama High School Graduation Exam 3rd Edition is aligned with 11th grade standards according to the Alabama State Department of Education and was first required for the graduating class of 2001 (Gayler, Chudowsky, Kober, & Hamilton, 2003).

By 2004 (ASDE, 2003) students had to must pass all five sub-tests in order to earn a high school diploma. There are five subject area tests included: science, mathematics, language, reading, and social studies. In 2004 Alabama was one of ten states that had a science subject area test as of 2004 according to (Gayler, Chudowsky, Kober, & Hamilton, 2004). Each subject test contains 100 multiple choice questions with the exception of the reading test, which contains 84 items (ASDE, 2003). According to Gayler and Kober (2004), Alabama and Tennessee are the only states that rely only on multiple choice questions tests.

As in most states Alabama students have additional opportunities to take the test if they fail (Gayler, Chudowsky, Kober, & Hamilton, 2003). However, Alabama students are given six opportunities to pass the graduation exam, more
than almost any other state (ASDE, 2003). All students receive a pass or fail grade. After taking the test, a student receives an individual student report broken down into questions passed by each domain area and each behavioral objective. Alabama students who fail the graduation exam receive remedial help from their local schools. Funding for this remediation is provided through the High Hopes Program (ASDE). According to Gayler et al. (2004), Alabama is one of eighteen states that required school districts to provide remediation for students struggling to pass the graduation exams. Alabama students are provided with remediation services until the age of 21 (ASDE, 2003).

Recently, the Alabama State Board of Education adopted a new five-year Alabama Student Assessment Plan (Alabama Education News, 2009). According to the Alabama Education News (AEN) the Alabama High School Graduation Examination (AHSGE) would be replaced with end-of-course exams by 2011-2012. “A student’s score on the final exam would be an embedded graduation requirement but not the sole determining factor in passing the class or in graduation,” according to Dr. Tommy Rice (AEN, p. 7). Alabama’s change to end-of-course exams will have several benefits. First, the end-of-course exams would eliminate 15 days of testing that are currently needed under the AHSGE system, thus increasing instructional time for students and teachers (AEN). Second, students would be taking the test right after they finish a course and would have less time to forget the material.

According to AEN, the state would also pay for all 11th graders to take the ACT one time. “This plan gives all students the opportunity to take the ACT,
including those who had never planned to take it or go to college. Under our ACT plan, those students may discover they have the scores and the potential to go to college” (2009, p. 7). The Alabama Student Assessment Plan also includes a revision of the ARMT in order to make it a comprehensive assessment (AEN). This would eliminate the need for Stanford 10 (grades 3-8); thereby reducing costs and increasing instructional time (AEN).

The accountability movement has also led to the development of several tests in the state of Mississippi. The Mississippi Curriculum Test, Second Edition (MCT 2) is a criterion-referenced test aligned with the Mississippi State course of study (Mississippi Department of Education, 2010a). According to the Mississippi Department of Education (MDE), the MCT 2 is made up of three subject area tests: reading, language, and mathematics. Students in grades 3 through 8 are required to take the MCT 2 tests (MDE). The MCT 2 tests were designed to ensure compliance with the NCLB Act (MDE). The data gathered by MCT 2 tests are used in the State Accountability System to help measure Average Yearly Progress (AYP) in grades 3 through 8 (MDE).

According to the Mississippi Department of Education (MDE), science assessments were developed for elementary and middle school grades to comply with the NCLB Act (2010b). The science assessments are criterion-referenced tests (CRTs) that are administered in grades 5 and 8 (MDE). The assessments are aligned with the Mississippi Curriculum Science Framework 2001 (MDE). The data gathered by these tests will be used to provide information the State Accountability model to ensure AYP (MDE).
The Mississippi Department of Education (MDE, 2010c) worked with the Pearson group to develop the Mississippi Writing Assessment (MWA). The MWA is administered in grades 4, 7, and 10 (MDE). The MWA was developed to help assess quality of writing of Mississippi’s students. The MWA requires students to respond to an expository prompt or a position prompt. Student responses are then graded using three grade-specific rubrics (MDE).

High school students in the state of Mississippi are assessed by the Mississippi Subject Area Testing Program (SATP) developed by the Mississippi Department of Education (MDE, 2005). According to MDE, the SATP helps ensure Mississippi’s compliance with the NCLB Act (2005). The SATP is a series of CRTs in four subject areas: Algebra I, Biology I, English II, and U.S. History. The SATP is aligned with the Mississippi State curriculum for each of the subject area tests (MDE). SATP scores are used in the state’s accountability model and to help schools measure progress toward their AYP goals (MDE).

Many states offer alternative routes to receive a diploma if students cannot pass the graduation exam. According to Gayler et al. (2004), 19 states offer some type of alternate assessment, exemption, waiver, or alternate diploma. The states of Virginia, New Mexico, Iowa, and Alabama offer students two diploma options for students who fail the state graduation exam (Baytops, McMahon, Padden, Walther-Thomas, & Vernon, 2003). Alabama has developed Adult Alternative High School Diplomas for those students who fail to pass the Alabama High School Graduation Exam (ASDE, 2003). Students are required to complete the basic Alabama high school coursework and pass the General
Educational Development test in order to receive the Adult Alternative High School Diploma (Daugherty, 2004).

Statement of the Problem

Another outcome of NCLB is an increase in information available to schools, parents, and government. School districts must develop and distribute a report card on the performance of each school. The report card must contain information about student performance on the assessments required by the school system and the state (Hanson, Burton, & Guam, 2006). The data must be broken down into the 9 student subgroups required by NCLB (Hursh, 2005). Clearly, communication of student performance data is increasing between government, schools, and parents (Hanson et al.). New assessments required under the NCLB Act will produce a massive amount of data for school systems to use in order to better analyze the success of their curriculum. Schools and districts will use this data to develop their school improvement plans in order to meet AYP.

A major concern is that teachers and administrators may not have adequate training to interpret the massive amount of data produced by these new tests (Hollenbeck, Tindal, & Almond, 1998; O'Sullivan & Chalnick, 1991). Hollenbeck et al. (1998) reported that fewer than 50% of the states even required a course in testing and measurement for teacher licensure. Hollenbeck et al. and Popham (2006a) asserted that teachers need sufficient training in order for test scores to be valid and interpreted in a correct way.
Schafer and Lissitz (1987) found measurement training was even less common in administrator training programs than teacher training programs. Impara (1993), in a study involving the National Association of Secondary Principals, found numerous deficiencies in the assessment literacy of the educational leaders in the United States. O'Sullivan and Chalnick (1991) found there was a lack of assessment training in principals and that assessment training was commonly not required for certification. Instructional leaders may lack the assessment skills necessary to interpret the data collected by new testing programs created as a result of the NCBL Act. Educational leaders may have difficulty determining which programs are effective in improving student performance if they lack the training to interpret the data and information produced by these measurements.

Since the NCLB Act has increased the importance of test data in the school improvement process, standardized testing has become the benchmark for measuring the success of students, teachers, and schools. Educators must be able to analyze the test data and develop a plan to improve the schools’ test scores. As a result, it has become very important that our education leaders are “assessment literate” (Paterno, 2001). Clearly, educators need sufficient training in testing, measurement, and statistical reasoning in order to analyze and make sense of this information. Popham (2003) stated, “All of us need to promote increased assessment literacy on the part of educational practitioners, educational policymakers, the public and especially parents of school age children” (p. 47). The problem under investigation in this study is: Do educators
have sufficient training in testing, measurement, assessment, and statistics to use these tests to predict future student achievement or are these tests created merely to comply with the NCLB Act?

Research Questions

1. Does method of certification make a difference in the training of school personnel in testing, measurement, assessment, and statistics based on method of certification or highest degree held?

2. How much training do teachers and administrators report having in testing, measurement, assessment, and statistics?

3. Can teachers and administrators identify the theoretical differences between norm-referenced and criterion-referenced tests?

4. Can teachers and administrators identify the major concepts of reliability and validity?

5. Can teachers and administrators identify potential misuses of assessment?

6. What are teacher’s attitudes on the application of educational statistics and their use toward education?

7. Will there be a difference in assessment literacy on norm-referenced/criterion referenced tests, validity/reliability, and potential misuses between teachers and administrators?

Purpose of the Study

The purpose of this study is to investigate the assessment literacy of teachers and administrators in regards to CRTs. First, the research will seek to
determine the amount of training in testing, measurement, assessment, and statistics that those teachers and administrators have received. Second, the researchers will investigate teachers’ and administrators’ knowledge of the theoretical basis of norm-referenced tests (NRTs). Third, the study will determine the participants’ knowledge of test validity and reliability. Fourth, the participants’ ability to recognize common misuse of test data will be investigated.

Hypotheses

1. There will be no significant difference in assessment literacy on norm-referenced/criterion referenced tests, validity/reliability, and potential misuses of school personnel based on degree.
2. There will be no significant difference in assessment literacy on norm-referenced/criterion referenced tests, validity/reliability, and potential misuses based on certification.
3. There will be no significant difference in assessment literacy on norm-referenced/criterion referenced tests, validity/reliability, and potential misuses based on years experience.

Definitions

1. Administrators: For the purpose of this study, administrators include principals, assistant principals, and administrative interns assigned to a school that is taking part in the study.
2. Arithmetic mean: For the purpose of this study, “The sum of all scores divided by the number of scores” (Coladarci, Cobb, Minium, & Clarke, 2004, p. 62).
3. Assessment: “The planned process of gathering and synthesizing information relevant to the purposes of a) discovering and documenting students’ strengths and weakness, b) planning and enhancing instruction, or c) evaluating and making decisions about students” (Cizek, 1997, p. 10).

4. Assessment literacy: For the purpose of this study, “The possession of knowledge about the basic principles of sound assessment practice, including terminology, the development and use of assessment methodologies and techniques, familiarity with standards of quality in assessment” (Paterno, 2001, p. 2).

5. Construct validity: The consistency between two theoretically-derived definitions of concepts or constructs. E.g. construct validity can be established when two tests that measure the same construct produce highly related scores (Zeller & Carmines, 1981).

6. Content validity: “The content validity of a test is the degree to which the items of that test are a representative sample of the content universe and or behavior of the domain assessed” (Zeller & Carmines, 1981, p. 387).

7. Criterion-related validity: Refers to the predictability of a future score when scores from a current measure are available (Quilter, 1999).

8. Criterion-Referenced test: “Criterion-referenced measures compare the student in relation to the level of performance he will be expected to achieve in a carefully defined domain of behaviors” (Popham, 1974, p. 254).


11. High Stake Testing: “Achievement tests being used 1) to make important decisions about students or 2) to evaluate the teachers who taught those students” (Popham, 2003, p. 45).

12. Instructional enhancement: For the purpose of this study, “Improving the quality of the educational experiences provided to students” (Popham, 1999, p. 13).

13. Mean: “The center of gravity of the distribution such that the weight of the score above the mean exactly balances the weight of the scores below it. Another way of looking at the mean is that the point of a distribution such that the algebraic sum of the differences of all the scores from the point is zero” (Shavelson, 1996, p. 93).


15. Norm-Referenced Test: Assessments, “designed to determine an individual’s relative standing in comparison with internal or external group” (Popham, 1974, p. 254).
16. Percentile: The score on a test below which a given percentage of scores fall (Frisbie, 2005). Example: The score of 85 on a science test is equal to or greater than 78% of the scores on that particular test. The student would have a percentile of 78.

17. Percentile rank: Most commonly the percentage of scores in a specified distribution that falls below the point at which a given score lies (Frisbe, 2005). Example: A test score which is higher than 80% of the scores would be the 80th percentile.

18. Reliability: “The consistency or dependability of a behavioral measurement. The notion is that assuming the subject is in a steady state, a measure on that subject should give exactly the same reading upon repeated measures with same instrument or with equivalent instruments that are used interchangeably to measure the same thing” (Shavelson, 1996, p. 473).

19. Test: “A systematic procedure for observing a person’s behavior and describing it with the aid a numerical scale or a category system” (Cronbach, 1970, p. 26). Examples: ACT, SAT, Iowa Basic Skills, or teacher made classroom measurement.

20. Test Misuse: When a test is used in manner that harms students and teacher or for a purpose that the test was not developed (Popham, 2003).

21. Standardized Achievement Test: “Any examination that is administered and scored in a standard, predetermined manner” (Popham, 2000, p. 12).
22. Standard deviation: “An average variability of scores in the distribution measured in units of the original score scale” (Shavelson, 1996, p. 82). A high standard deviation indicates the scores are spread out over a large range from the mean. A low standard deviation indicates that the data points tend to be close to the mean.

23. Validity: Considered an evaluative judgment about the degree to which test scores are appropriate for making certain educational decisions (Messick, 1994).

24. Variability (of a distribution): “Describes the spread or range of scores in the distribution” (Shavelson, 1996, p. 82).


Delimitations

The participants in this study were practicing educators that included teachers, principals, assistant principals, and counselors within the coastal areas of the states of Alabama and Mississippi. Any inferences in the findings of this study to a larger population should keep in mind this limitation.

Assumptions

The participants of this study were representative of educators across the coastal area of Alabama and Mississippi. Educators answered the literacy questionnaire honestly and to the best of their ability.
Justification

The ‘No Child Left Behind’ Act requires teachers and administrators to analyze test data and develop school improvement plans based on this information. “Teachers and school administrators in the current NCLB era are expected to have a sophisticated understanding of test results to use them to make data-based decisions about classroom instruction, and to communicate them to others” (Zwick, Sklar, Obsipo, Wakefield, Hamilton, Norman, & Folsom, 2008). Teachers without a thorough knowledge of assessment are a clear liability for any school or school district (Popham, 2000). Popham (2006d) stated, “Today’s educational leaders need to understand the basics of assessment or they are likely to become yesterday’s educational leaders” (p. 13).

Assessment literate individuals are more likely to know the difference between good and poor assessment practices. Assessment literate people understand the potential negative impact and misuse of inaccurate assessments (Stiggins, 1995). Stiggins (1990) stated, “The assessment community knows that teachers and administrators are in desperate need of assessment training; teachers and administrators know they need assessment training, but policy makers often are not aware of this gap in professional preparation” (p. 95). If teachers and administrators do not have adequate training in testing, measurement, and statistics, universities may need to provide more training to adequately prepare educators in these areas.

Carter (1984) recommended that preservice coursework in measurement and assessment be examined to see if a relationship existed with teacher
assessment knowledge. Subsequent studies have found that teachers may lack the training to be assessment literate (Hollenbeck, Tindal, & Almond, 1998; O’Sullivan & Chalnick, 1991; Plake, 1993; Popham, 2006b; Stiggins, 1995). Schafer and Lissitz (1987) found that less than half of the teacher education programs require one course in assessment, measurement, and statistics for completion. The situation is even worse for administrator programs. Assessment has become more important since the passage of NCLB. Has the amount of training required by colleges increased since the passage of NCLB? A 2004 study found no increase in testing, measurement, and statistics training (Lukin, Bandalos, Eckhout, & Mickelson, 2004). Furthermore, recent studies have started to link student performance with their teachers’ knowledge of assessment. Stiggins (1999a) reported that increasing teachers’ assessment knowledge could cause a 0.4 to 0.7 standard deviation increase in students’ standardized test scores. Clearly, anything that could increase student performance so dramatically should be examined by the research community.

This study will investigate assessment literacy of teachers and administrators and help to identify areas that teachers need for professional development to become more competent in the use of assessment. Teachers could use the instruments from this study to help identify their own needs for professional development. This research could also help to determine future needs in assessment training for teachers and administrators in college and other certification programs.
CHAPTER II
REVIEW OF RELATED LITERATURE

History of Large Scale Assessments

Large scale assessment has a long history. According to Gerberich (1963), one of the earliest tests recorded was the Old Testament of the Bible. The Jephthah Test was an oral exam to determine if a person could pronounce the word “Shibboleth” (Gerberich). Jephthah was one of the twelve judges of the Old Testament, Judges, 10-12 (New International Version). The test was used to determine the Gileadites, which could pronounce the “sh” sound, from the Ephraimites, that could not pronounce the sound (Popham, 1990). Popham reported that everyone who failed the test was beheaded. The entire story of the Jephthah Test can be found in Judges, 12:1-7 (New International Version).

In China, as early as 2200 B.C., civil servants were required to pass a competitive examination and interview process in order to obtain and keep their jobs (Popham, 1990). The candidates had to pass a competency test in music, writing, archery, arithmetic, and horsemanship (Popham). These early competency tests had a very high failure rate of 93 to 97 percent (Popham).

European contacts with China led to the development of civil service examinations in Britain and the United States (Popham). The British established a civil service examination in the 1850’s. In the United States, President Grant set up the Civil Service Board in 1871.

Universities also had a long history of the use of large scale assessments. Schools in Greece used assessments to determine the proficiency of students in
language and fine arts (Gerberich, 1963). The University of Bologna, as of 1219 A.D., required oral law exams (Gerberich). According to Popham (1990), the Louvain University System in 1540 made extensive use of written examinations when it established the Jesuit Order. Cambridge and Oxford required written examination in the early 1800's (Popham).

According to Mathews (2006), Horace Mann was advocating standardized testing in the United States as early as 1845. In 1845 written exams were required in Boston Public Schools and Harvard started its first entrance exams in 1851, according to (Michigan State University, 2009).

Subsequently, Alfred Binet developed the first intelligence test in 1905 (Matthews, 2006). This new intelligence test was developed from thirty different tests (Popham). Binet was able to fuse these individual tests into one comprehensive measure of intelligence. Alfred Binet’s work provided the framework for modern IQ tests. These intelligence tests led to the concept of normal distribution or Bell-shaped Curve. The recognition that human intelligence was normally distributed within the population helped in the development of NRTs.

The potential of NRTs to assess differences in abilities of the individual was quickly recognized by the military. The necessity to test individuals quickly led to the creation of the multiple-choice test which was created in 1914 by Fredrick J. Kelly (Mathews). Classical test theories were used to determine differences in individuals in 1918, during the First World War (Popham, 2000; Wineburg, 2004). Popham stated, “The Alpha was an intelligence test measuring
an Army recruit’s aptitude for success in an officer training program” (p.12). During World War I alone, more than 1.7 million men were tested using the Army Alpha Test (Popham, 1990; Wineburg, 2004). Military intelligence tests have been given to millions of recruits over the past century. These tests have produced a tremendous amount of data that had great influence on our classical test theories.

During the 1920’s and 1930’s there were three major developments in the area of standardized testing. First, the Roschach’s Inkblot Test for personality measurement was developed in 1921 (Gerberich, 1963). Secondly, in 1926, the multiple choice SAT was developed and implemented for college applicants (Mathews, 2006). Third, Harthshorne and May, in 1930, developed techniques to measure cooperativeness persistence and honesty in different situations (Gerberich). The military quickly recognized the potential use of these tests. The Office of Strategic Services used these new situational tests to select spies for the United States government during the Second World War (Gerberich).

Since the 1950’s national and international events have effected assessment in the United States. In 1957 the launch of Sputnik shocked the United States and led to increased scrutiny of the country’s public school systems (Mathews, 2006). As a result, the National Defense Education Act of 1958 funded increased training in Math and Science education. The Act also expanded and increased the development of testing to measure students’ achievement.
The 1960’s saw several important developments in the areas of assessment. In 1965, the Elementary and Secondary Education Act was passed (Michigan State University, 2009). The Act required any school receiving federal money to show that educational goals were being met. Standardized tests were acceptable proof of reaching these goals. The National Assessment of Educational Progress (NAEP) was meant to measure the progress of the United States Public Schools by using a National Standard (Michigan State University, 2009).

In 1983, “A Nation at Risk,” a report prepared by a national commission, stated that educational standards across the United States were too low (Mathews, 2006). As a result, in 1988 the National Assessment Governing Board was created to oversee testing in the United States (Mathews). The National Assessment Governing Board developed new standards for the NAEP testing program.

With the passing of No Child Left Behind Act (NCLB) in 2001 states and local school districts became required to develop plans for continuous improvement until all students meet proficiency. Large scale assessments have become more important because of the Annual yearly progress (AYP) requirement of NCLB. AYP plans must be submitted by each school district to show how the individual districts are working toward proficiency (Hanson et al., 2006). AYP must be met by each school and district. Each individual student and the school population as a whole, as well as various subgroups, such as: race, gender, social economic status, and special education, must show annual
improvement (Hursh, 2005). The law mandates that schools be accountable for the success of every child. The law also mandates that 95% of all students, including those with special needs, participate in the assessment in order to meet the AYP goal for participation (Hursh). For the first time, a federal law (NCLB) mandates that schools are accountable for the success of all students.

Types of Large Scale Assessments

Since NCLB, testing has become an important element of the educational process in the United States. The current educational system in the United States generally uses two basic types of tests to evaluate students: NRTs and CRTs. NRTs were designed to determine a student’s knowledge in relation to a group of students on the same test instrument (Popham & Husek, 1969). On the other hand, CRTs were designed to assess a student’s mastery of clearly defined competency of behavioral objectives in a subject area (Linn & Gronlund, 2000; Popham & Husek; Symth, 2008). According to Goodstein (1982), the main difference between NRTs and CRTs is in the manner in which the scores are interpreted. Since the major difference in the two tests is in interpretation, the CRTs offer no real advantage over traditional NRTs (Popham, 1978). Both types of tests can provide meaningful information to educators and parents.

According to the American Psychological Association and National Council on Measurement in the Standards for educational and psychological testing (1999), “Norm-referenced test interpretations are based on a comparison of a test taker’s performance to that of other people in a specified reference population” (p. 178). The NRTs helped develop understanding of the normal
distribution which led to the development of modern statistical techniques. In
addition, NRTs heavily influenced the development of educational statistical
reasoning.

Most teachers are very familiar with the NRT (Popham, 1978). Also,
NRTs provide educators with data that are familiar, such as: means, modes,
standard deviations, and percentile ranks. According to Popham and Husek
(1969), norm-referenced tests are useful when there is a need to compare an
individual to a group. Achievement tests are norm-referenced tests; examples
include: ACT, SAT, and Stanford.

Theoretical Framework of Large Scale Assessment

Norm-Referenced Tests

NRTs were developed from the pragmatist educational philosophy
(Terwilliger, 1977). Terwilliger asserted that pragmatist philosophic ideas led the
development of conventional approaches to assessment and stated:

The pragmatist is primarily concerned with practical choices and
consequences of such choices. Because both the choices and their
consequences can be tied to individual differences in abilities and skills,
differentiation among individuals is deemed desirable. A grading system
which is defined with reference to the individual is the optimal system.

This is what is called the norm-referenced point of view. (p. 26)

Classical test theories were developed and studied for over 150 years
(Traub, 1997). NRTs are firmly grounded in classical test theories (Popham &
Husek, 1969). Since these tests are supported by classical test framework, there
is a general agreement on how the tests should be constructed and interpreted (Popham & Husek). Furthermore, methods for calculating test validity and reliability have been determined by the academic community. The reliability coefficients, as well as other major indexes, (such as mean, median, mode, and standard deviation) are calculated based on variability around the mean (Livingston, 1972). It is important to remember that NRTs are based on variance of score from the mean of the population (Popham & Husek). Thus, variability is a key component in classical test theory. However, the validity and reliability of any test can be compromised by the poor interpretation of these scores (Frisbe, 2005; Popham, 2006b; Smyth, 2008).

Criterion-Referenced Tests

The developments of CRTs are deeply rooted in the Behaviorist Learning Theory in Psychology. The works of Thondike, Hull, Skinner, Bloom, and Gagne have all aided in the development of criterion-referenced assessment. According to Terwilliger (1977), B. F. Skinner’s ideas have heavily influenced educational learning theories. Terwilliger stated that the “Skinnerian model includes: 1) the precise objectives of instruction, 2) the exact instructional sequences to be employed, and 3) the specification of the criteria used for judging whether the objectives have been attained” (p. 23). The Skinnerian learning theories led to the mastery approach of evaluation and instruction developed by Benjamin Bloom (Terwilliger). Benjamin Bloom became the father of instructional objective writing for mastery learning. In addition, Mayo (1974) summarized the steps in the mastery approach for learning. First, students should learn in a cooperative,
not competitive, setting. Second, standards for mastery of the lesson should be set in advance of the lesson. Third, formative assessments should be used with each lesson. Fourth, remediation should be provided for students that fail to achieve initial mastery of the lesson. Fifth, students should be given additional time to master objectives if needed.

Noted Behaviorist B. F. Skinner (1954) stated:

The whole process of becoming competent in any field must be divided into a very large number of very small steps, and reinforcement must be contingent upon the accomplishment of each step. This solution to the problem of creating a complex repertoire of behavior also solves the problem of maintaining the behavior in strength. By making each successive step as small as possible, the frequency of reinforcement can be raised to a maximum while the possibility of adverse consequences of being wrong is reduced to a minimum. (p. 94)

Edward Thorndike’s works led to the development of the objective test and he is considered the father of scientific measurement (Shepard, 2000). Benjamin Bloom developed the idea of writing instructional objectives to set the standard for mastery learning for students. These student learning objectives could then be used to help develop a test to assess student mastery of an objective. Terwilliger (1977) stated that the advocates of mastery learning approach are primarily concerned with whether students master a criterion rather than individual differences among students. According to Terwilliger, “The first preference of the behaviorists is not to employ differential grades at all, but if
there is no choice in the matter, he/she will prefer to base grades upon pre-determined performance criteria. This is what is commonly called the criterion-referenced point of view” (p. 24).

NCLB has mandated the success of all students by the year 2014. The mastery approach to learning and criterion assessment may provide a possible avenue for this to take place. In 1971 Benjamin Bloom seemed to foresee this thought when he stated:

The complexity of skills required by a work force of any highly developed nation like the United States suggests we can no longer assume that completion of secondary and advanced education is for the few. Investment in human resources through education has a greater return rather than capital investment. We cannot return to an economy in which educational opportunities are scarce, but rather must provide enough opportunities that the largest possible proportion of students will acquire the skills and knowledge necessary to sustain the society’s growth. (p.48)

Thus, the most important investment a society can make is in the education of its citizens.

Glaser (1963) was the first person to use the term CRT when he stated: Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual’s achievement level falls at some point on the continuum as indicated by the behaviors he displays during testing.
The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measurements of achievement or proficiency. (p. 519)

Glaser continued by stating, “Measures which assess student achievement in terms of a criterion standard thus provides information as to the degree of competence attained by a particular student which is independent of reference to the performance of others” (p. 519). Thus, the main difference in NRT and CRT is how they are scored and interpreted.

The main difference between NRT and CRTs is the standards by which the students’ performance is judged (Shavelson, Block, & Revitch, 1972). In a NRT, the student’s performance is gauged against the mean. CRTs are based on variability around a cut-score (Livingston, 1972). These tests are interpreted based on student test performance in relationship to a pre-set cut-score (Livingston). They are not based on variability from the mean (Popham & Husk, 1969). Furthermore, Popham and Husek stated that, “Variability is irrelevant in criterion-referenced tests” (p. 3). Reliability coefficient and other basic statistics are calculated based on their variability around the test’s cut-score (Livingston). Variability around a mean is the foundation for classical test theories. Thus, CRTs have been developed as an outgrowth of our classical test framework.

CRTs are based on the idea that learning occurs within a cognitive domain (Goodstein, 1982). Cognitive domains may be broken down into instructional objectives or student learning outcomes. Popham and Husek (1969) indicated that criterion measurements may be used to determine student mastery of
instructional objectives. Thus, criterion measurements could be used to help determine the effectiveness of a curriculum or refine the learning process (Behuniak & Tucker, 1992). Teachers are using more CRTs, hoping they will be more useful in measuring educational outcomes (Popham, 1978). Thus, a well constructed test can directly measure how well the teachers are teaching and the students are mastering the curriculum. Also, CRTs may be more useful to local educators than NRTs because they coincide with the local curriculum (Popham, 1977).

Popham and Husek (1969) recognized that tests based on classical test theories were not useful measurements of student learning in many situations. This sentiment was echoed by Hart and Sciutto (1996) when they determined that there were limitations of traditional tests to measure student achievement. The comparative nature of traditional tests inhibits their usefulness in assessing schools or teachers (Popham, 2003). CRTs may help bridge a gap between traditional test theory and useful test practices. According to Foegen and Deno (2001), the potentials of CRTs to measure student achievement are high and recommended further research to help refine these measurements for use at the high school and middle school level.

CRTs are created with a clearly defined domain of behavior to be assessed (Popham, 1977). CRTs can be constructed to be aligned with the local curriculum of a school district (Evans, 1975). Thus, a good CRT will provide information on classroom instruction related to the behavioral objectives on the test (Evans). A well constructed test can directly measure how well the teachers
are teaching and the students are mastering the curriculum. CRTs may be more useful to local educators that norm-referenced tests because they coincide with the local curriculum (Popham, 1977).

CRTs that assess mastery or non-mastery of behavioral objectives must be constructed very carefully, and should clearly relate to the curriculum. A good curriculum should include well defined, clearly stated objectives (Evans, 1975). The learning objectives should be written in terms of the measurable behaviors that are desired from the learner (Popham, 1978). A well developed CRT will measure what the students have mastered certain learning objectives. Test questions that are designed to measure whether students have mastered these learning objectives are called test specification, according to Popham.

Shavelson et al. (1972) recommended several steps in developing quality CRTs. The developer must determine how they will calculate: reliability, validity, and other basic statistics. The cut-score needed for student mastery of the instructional domain should be predetermined. Shavelson et al. recommended that students should be required to answer 75 percent of the questions for mastery. How the test will be used in the instructional process should also be worked out before the test is implemented.

Another important attribute of any well developed test is the number of test items. Evans (1975) stated that a tryout test program should be used to help determine the minimum number of test questions per each behavioral objective. Popham (1978) stated that 2-4 items per behavioral objective are typical. Sometimes a CRT will assess competency in a domain with one test item. This
should be considered a flaw in developing a CRT to assess mastery of any learning objective. Popham stated that more test items per objective should be used in particular high stakes testing such as the high school graduation exams. The test should be developed with the realization that the whole curriculum must be represented in the test (Evans). According to Popham, a well-developed CRT will focus on a smaller number of behavioral objectives such as six to 12, rather than a large number.

A well developed test should have a thorough validity appraisal; especially in regard to the objectives it measures (Evans, 1975). Valid tests are supposed to measure the knowledge and skills the student should possess after the curriculum has been taught (Davis, 1998; Odom & Morrow, 2006). Generally, there are two broad types of validity: descriptive validity and construct validity. Descriptive validity can be thought of as to what extent the test questions are aligned to the test’s descriptive scheme (Popham). Construct validity is generally defined as the extent to which the test performance is related to the content or curriculum the test is intended to evaluate (Evans). Descriptive and constructive validity are both needed in order to have a valid test or measurement of performance. Validity cannot exist without the test having a high degree of reliability (Goodstein, 1982). Validation considerations should not be overlooked when developing any test (Popham).

Good tests demonstrate a high degree of reliability. Test reliability is basically whether students would perform the same on the test if the test was given on more than one occasion (Davis, 1998). Another way of thinking about
reliability is whether students would earn the same scores on the same test items on different occasions (Davis). Reliability could also be viewed as whether students would receive similar scores on a set of test items assessed by different questions (Davis). It is important to remember that reliability, as defined above, cannot be measured in some form of testing situation. Goodstein (1982) stated that reliability of CRTs can be considered in three areas: first, the reliability of the classification of behavioral objectives by domain, second, the reliability of the criterion-referenced test scores, and third, reliability in the domain scores of each domain. CRTs can be used for testing instructional objectives, individual competence, and student placement (Goodstein; Linn et al., 2000; Smyth, 2008).

Tests with reliability coefficients of .80 to .90 are generally considered to have good reliability (Odom & Morrow, 2006; Popham, 1978). Such tests are generally 50 to 100 questions (Popham). CRTs, however, are generally shorter in length, testing six to 12 objectives with 10 to 15 questions per domain (Popham). The lower number of test items results in a lower reliability coefficient than educators are used to (Popham). Popham stated that reliability should be reported for each sub-area/behavioral domain in a CRT. It is important to note that a test cannot be considered valid if reliability cannot be demonstrated (Goodstein, 1982; Odom & Morrow, 2006). After reliability and validity have been established for the test, the pass/fail score must be established using comparative data provided by the test results.

It is important to understand that reliability estimates may be manipulated by simply adjusting the cut-score (Livingston, 1972). Test issues such as validity
and reliability have not been agreed upon in regards to CRTs. Goodstein (1982) stated that reliability in CRTs can be looked at in three areas: reliability of the classification of behavioral objectives by domain, the reliability of the CRT scores, and reliability in the domain scores of each domain. Shavelson et al. (1972) recommended that reliability should be calculated for each subscale on the test, “The subscale reliability will be lower than the total score reliability” (p. 136). It is important to remember that criterion scores are only valid and reliable if they meet the major assumptions of our classical test theories (Livingston). Criterion-referenced tests can be used for testing instructional objectives, individual competence, and student placement (Goodstein).

Interpretation of comparative data is an important part of a CRT. Comparative data is information on how well the students perform in relation to each other within an educational domain and on various behavioral objectives. The comparative data is used to help develop the cut and passing score for CRTs (Evans, 1975). The question arises as to how high students must score to achieve proficiency? According to Dilendik (2001) and Smyth (2008), criterion levels and cut-scores are the basic level of competency required of the student and should be determined before the tests are administered. Typical CRTs may have 75% to 90% high proficiency level as opposed to only 5% in a NRT (Dilendik). CRT performance is based on mastery, not normative data like a NRT (Popham, 1978). Traditional evaluation systems are based on normative data (Dilendik). NRTs were designed to rank students in comparison to larger groups (Popham, 1977).
According to Wise, Lukin, and Roos (1991), the use of criterion-referenced testing has increased throughout educational systems in the United States. They further stated that the skills necessary to successfully implement such testing were different than traditional testing. They recommended that additional training in assessment was essential to efficiently use CRTs. This sentiment was echoed by Frisbie (2005) when he recognized that many teachers failed to make the transition from NRTs to CRTs. Frisbie noted that classical indices were calculated in a similar fashion. However, standards for use and acceptability were different between the two types of tests. Several studies (Wise et al., 1991; Frisbie, 2005) recognized that teachers would need additional training to effectively use criterion-referenced measurements. Evidently, many researchers in the assessment community were starting to recognize that additional training in measurement would be needed by educators to effectively do their job.

Assessment Literacy of Teachers and Administrators

Researchers have found that teachers spend a lot of time on assessment. Stiggins (1988) stated that teachers spend one-third of their time on assessment related activities. In a study conducted in 1991, before the passage of NCLB, Stiggins determined that assessment took as much as 50% of a teacher’s time. Assessment related activities could include developing and grading: tests, quizzes, lab reports, and research papers. Clearly, teachers spend a lot of time on assessment activities. Since the passage of NCLB, the number of standardized assessments given to students has increased. Therefore,
teachers, administrators, and students are conceivably spending even more time on assessment.

Assessment is a major part of education in the United States educational system. However, questions have been raised about the level of assessment literacy of teachers. Stiggins (1988) noted that teachers are neither trained nor prepared to perform their role in the assessment process. Wise et al. (1991) found that 47% of teachers reported that their assessment training was not adequate and that most of their assessment knowledge was a result of trial and error.

Many teachers report that they are not prepared sufficiently in the area of assessment. Perhaps an analysis of teacher training in measurement will shed some light on this concern. Many states do not require assessment training to be for certification as a teacher (O’Sullivan & Chalnick, 1991). Hills (1991) found that only four states actually required teachers to take a measurement course. Stiggins (1991) also found that 25 states had developed some type of competence standard in the area of assessment. Ten states including AL, AK, AZ, CA, IA, MT, ND, TX, WI, and WY actually required teachers to take at least one course in assessment in order to be certified (Stiggins). This seems to support Wolmut’s (1988) earlier findings that only 20% of the states required a measurement or assessment course for teacher certification. These fifteen states, CO, CT, DE, FL, HI, IN, NY, OH, OK, OR, TX, UT, VT, VA, and WA, have developed teacher certification standards for competency in assessment (Stiggins). However, these states did not require formal coursework in the area
of measurement assessment. Fifty percent of the states have not developed any standards or course requirements for assessment literacy of teachers. To date, there is no certification exam for assessment literacy of teachers in any state (Zwick et al., 2008). Evidently, many states have not addressed competency in assessment as a condition for being licensed to teach (Stiggins, 2002). Furthermore, Stiggins (1999b) stated, “We have failed to impose licensing standards that require teachers to be competent in assessment in order to practice” (p. 195).

The situation for school administrators appears to be even worse. To date, no state requires school administrators to be assessment literate in order to be certified (Zwick et al., 2008). In 2005, Adams and Copeland found that no state’s certification agency had defined the knowledge and skills needed by principals or school administrators to be assessment literate. State licensing policies seem to ignore the need to set standards for assessment literacy for school administrators. How can school administrators evaluate the data gathered under the NCLB if they lack assessment skills necessary to analyze the data?

The training provided to teachers and school administrators by universities and colleges seems to be a little better than what is provided by the state’s certification agencies. Roeder (1972) reported that 57.7% of institutions did not require a teacher to complete a course in measurement or assessment. Similarly, Schafer (1991) found that only about half of the education programs required in assessment or measurement course in order to graduate. Roeder
found only 12 % of the universities required one 3 hour class in measurement; 1.4 % required two or more classes in assessment. Often, assessment and measurement content required in teacher education programs has been embedded in other classes and not taught as discrete courses (Wise & Lukin, 1993). Measurement classes are often taught by instructors that are not measurement experts (Quilter, 1999). Noll (1955) examined college catalogs to determine the availability of assessment and measurement courses. Noll found that 82.5 % of colleges offered a course in measurement at the undergraduate level. However, only 21.2 % required a course for graduation. According to Schafer (1991), there was no major increase in assessment coursework requirements for teachers between 1960 and 1991. Perhaps this is why Stiggins (1991) stated that colleges have failed to make their education graduates assessment literate.

Administrator training programs require less coursework in measurement than teacher programs (Schafer & Lissitz, 1987). Schafer and Lissitz found that only 15 % of administrator training programs require a course in assessment or measurement. Master and doctoral programs in administration seem to require more coursework in measurement (Schafer & Lissitz). Clearly, administrator training programs may not require enough coursework, assessment, measurement, and statistics in order to be assessment literate. According to Jennings, The Center on Education Policy stated, “school leaders must be adept at using data to improve teaching and learning. The NCLB expects administrators to have a sophisticated understanding of assessment and to use
test results to make data driven decisions” (Jennings, 2002, as cited in Zwick, et al. 2008). How can school administrators develop a sophisticated understanding of assessment when they are not required to take coursework in this area? Stiggins (1999b) stated, “We have failed to prepare principals to provide proper supervision and leadership in assessment” (p. 195).

As stated earlier, even when colleges and universities require coursework in measurement, assessment, and statistics the classes are often taught by professors that lack expertise in measurement (Quilter, 1999). According to Schafer and Lissitz (1987), the competency of a professor in a specialized area is commonly measured by his/her ability to conduct research in that field. Schafer and Lissitz’s survey on research conducted by professors of measurement classes showed an average of the number of articles published by each instructor of measurement classes, “1.5 at bachelor-level institutions, 2.0 at master-level institutions, and 4.1 at doctoral-level institutions” (p. 61). Clearly, many professors of measurement classes are not conducting research in measurement. The competency of many professors of measurement classes could be in question based on this information.

Zwick et al. (2008) found that teachers and administrators have substantial gaps in their knowledge of statistics and assessments. Zwick’s study found that most teachers and administrators did not understand statistical concepts such as: z-score, standard deviation, grand mean, measurement error, and reliability. Stiggins (1998) also found that teachers often ignored the statistical aspects of assessment. Evidently, many teachers may not have a
good grasp of statistical techniques and their relationship to assessment
(Gullickson, 1986). Gullickson also found in 1986 they did not employ statistical
techniques to analyze the tests they employ. Teachers may not use statistics in
their assessment practices. According to Gullickson, teachers perceived
statistical analysis of test data as more work than the analysis is worth. This is in
direct contrast to fact that measurement experts have established a positive
impact on learning when a teacher uses a statistical analysis of test data
(Gullickson). Furthermore, teachers and administrators seem to be at odds with
what measurement experts deem appropriate (Lai & Waltman, 2008).
Educational practitioners seem to have a strong dislike for the statistical analysis
of test data (Mayo, 1967). Mayo reported that teachers loathe statistical
concepts and analysis. Perhaps this bias is caused by the fact that
administrators and teachers did not have adequate training in assessment,
measurement, and statistics.

The lack of assessment literacy in the education profession triggered the
action of Professional Education Organizations. In 1989, the Boros Institute of
Mental Measurement held a symposium to address the “crisis in teacher training
in measurement and assessment” (Luckin et al., 2004, p. 27). In 1990, the
National Council Measurement in Education (NCME), National Education
Association (NEA), and American Federation of Teachers (AFT) worked together
to set and publish Standards for Teacher Competence in Educational
Assessment (Plake, Impara, & Fager, 1993a). The American Federation of
Teachers, National Council on Measurement in Education, and National
Education Association (1990) developed seven principles for assessment literacy of teachers:

- **Standard 1** – Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
- **Standard 2** – Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
- **Standard 3** – The teacher should be skilled in administering, scoring, and interpreting - the results of both externally produced and teacher-produced assessment methods.
- **Standard 4** – Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
- **Standard 5** – Teachers should be skilled in developing valid pupil grading procedures that use pupil assessments.
- **Standard 6** – Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators,
- **Standard 7** – Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

After National Standards for Assessment Literacy were established, steps were taken to assess teacher knowledge of assessment based on these standards. Plake (1993) developed the Teacher Assessment Literacy Questionnaire as an instrument to assess knowledge aligned with these
assessment literacy standards. The questionnaire contained 35 questions, five per standard. Plake found that teachers scored highest on standard three: administering, scoring, and interpreting the results of assessment. Teachers scored the lowest on standard six: communicating assessment results standard six. However, Plake reported several other interesting findings in her study. Plake found the five questions were answered correctly 30% of the time. Only 13% of the teachers answered the question on reliability correctly. Teachers also had trouble answering questions on unethical or illegal practices in assessment (Plake, 1993). Plake’s findings seemed to indicate that teachers had trouble understanding validity, reliability, and unethical and illegal practices in assessment.

Educational leadership organizations have also studied the assessment literacy of school administrators. Impara (1993) reported on a study sponsored by the National Association of Elementary School Principals and the National Association of Secondary School Principals which found major deficiencies in assessment literacy of the school administrators in the United States. The study highlighted the need for National Standards for assessment literacy for school administrators.

Subsequently, the National Policy Board for Educational Administration developed 12 competencies for assessment literacy of school administrators (Ramirez, 1999):

- Understand the attributes and applications of sound student assessment
• Understand the attributes and applications of a sound school assessment system

• Understand issues involving unethical and inappropriate use of assessment information and ways to protect students and staff from misuses

• Understand assessment policies and regulations that contribute to the development and sound use of assessments at all levels

• Set goals with staff for integrating assessment into instruction and assist teachers in achieving these goals

• Evaluate teachers’ classroom assessment competencies and build such evaluations into the supervision process

• Plan and present staff development experiences that contribute to the use of sound assessment at all levels

• Use assessment results for building-level instructional improvement

• Accurately analyze and interpret building-level assessment information

• Act on assessment information

• Create conditions for the appropriate use of achievement information

• Communicate effectively with members of the school community about assessment results and their relationship to instruction. (p. 205)

The National Policy Board for Educational Administration (2008) revised the educational leadership policy standards for educational administrators in the United States. The policy standards were designed around building a better vision for educational leadership in the United States. The new standards have
an embedded component that stresses the need for assessment literacy for administrators:

- Standard 4: An education leader promotes the success of every student by collaborating with faculty and community members, responding to diverse community interests and needs, and mobilizing community resources. Functions: A. Collect and analyze data and information pertinent to the educational environment, B. Promote understanding, appreciation, and use of the community’s diverse cultural, social, and intellectual resources, C. Build and sustain positive relationships with families and caregivers, and D. Build and sustain productive relationships with community partners. (p. 14)

Importance of Assessment Literacy

Assessment literate individuals are more likely to know the difference between good and poor assessment practices. They understand the potential negative impact and misuse of inaccurate assessment (Stiggins, 1995). Assessment literacy is important for three main reasons. First, teacher assessment literacy has been linked to student success on large-scale assessment tests (Black & William, 1998). Second, a person’s assessment knowledge affects the validity and reliability of the test (Popham, 2006b). Third, lack of understanding assessment leads to misuse of assessments (Popham).

Teacher assessment literacy has been linked with student performance on standardized tests. Carter (1984) noted that teachers’ test making skills were
linked to students’ test taking skills. In a subsequent study, Black and William (1998) found a strong link between improving classroom and teacher assessment literacy and improving student standardized test scores. Black and William found that standardized test scores could be improved by as much as 15-20 percentile points, three-fourths of standard deviation, or four grade levels by increasing the quality of assessment practices in a school. This data was supported by similar studies that found students’ standardized test scores increased by .75 to 1.5 of a standard deviation when classroom assessment practices were improved (Meisels, Atkins-Burnett, Xue, & Bickel, 2003; Rodriguez, 2004). Interestingly, low performing students experienced the largest gain in standardized scores in all these studies. Such studies support the idea that increasing teacher and administrator assessment literacy can lead to dramatic increases in student achievement.

Some suggest assessment literacy of teachers and administrators can affect the validity and reliability of test scores. Carter (1984) found that lack of teacher assessment skills called into question the validity of test results. Lai and Waltman (2008) reported that, “teacher instructional practices may invalidate test scores used for accountability purposes” (p. 30). In particular, test preparation activities may cause problems with test validity (Lai & Waltman). Lai and Waltman found that inappropriate test preparation and administration caused the construct-irrelevant variance to increase in the scores on students. This increase in the construct-irrelevant variance created a systematic error as opposed to random error. An increase in systematic error jeopardizes the validity and
reliability of test scores. Hollenback et al. (1998) reported that, “The total error variance is negatively correlated to a test’s reliability” (p. 177). Thus, increasing test error results in decreasing the reliability and validity of the instrument and the scores collected using it. Lai and Waltman asserted that increasing the assessment literacy of the teachers can increase the validity of assessments by decreasing error in measurements.

Nolen, Haladyn, and Haas (1992) contended that, “Variation in test administration results in a lack of standardization fundamental to valid score interpretation and use” (p. 13). A common source of test variation can be caused by teachers failing to completely follow directions when administering tests to students. Test administrators increase test variation when they fail to supervise teachers to insure that all testing procedures are followed so that all students take the test under the same conditions. Frisbe (2005) asserted that incomplete directions for test administration were a major threat to the validity of test scores. This argument was echoed by Hollenbeck et al. (1998) who asserted that modification of testing procedures was a threat to validity and reliability of tests. Hollenbeck et al. also noted that many teachers do not realize that test preparation practices and inappropriate accommodations can make scores invalid, teachers failed to understand how to implement accommodations for special education students, and they have problems with test modifications for special education students during standardized testing. Evidently, test administration practices can raise or lower test scores and cause an inaccurate
measure of a student’s knowledge. Thus, poor test administration practices can reduce the reliability and validity of test scores.

Teachers seem to have trouble in communicating test results and may have trouble interpreting some types of test scores (Plake et al., 1993a). Teachers must be careful in how they interpret test scores or the validity of the scores can be compromised (Popham, 2006a). Hollenbeck et al. (1998) stated, “High academic standards are only as meaningful as the reliability and validity of the score’s interpretation” (p. 181). Thus, teachers and administrators need to understand how to interpret and communicate test results to students, parents, and other stakeholders (Popham, 2006c).

Misuses of Assessment

Using Student Test Scores to Evaluate Teachers

Assessment-illiterate teachers and administrators often misuse test data in a way that is harmful to the educational process. Popham (2003) reported that public schools in the United States used traditionally constructed large scale assessment tests. These tests were often being misused to evaluate teachers and schools and to make important decisions about students based on a single test score. Traditionally constructed tests are generally considered norm-referenced tests. However, Popham reported new standard-based criterion tests are being misused in the same way. This misuse of large scale assessment is occurring in almost every state (Popham). The misuse of test data lowers the instructional quality of the education that children receive.
Standardized achievement tests should not be used to evaluate schools or teachers for several reasons. Such tests are designed to measure student achievement not teacher or school achievement. According to the *Standards for Educational and Psychological Testing*, Standard 1.4, test scores can only be used for the purpose for which the validity and reliability have been established (Popham, 2003). Test scores must be valid and reliable to have any meaning. If administrators want to use these student test scores to evaluate teachers, they need to collect data and establish validity and reliability for this purpose; however, this work has simply not been done (Popham, 2003). Popham (2001) reported, “Some educational policy makers are advocating a teacher evaluation model that simply subtracts last year’s students’ average test scores” (p. 28). However, the students that teachers receive each year vary considerably in their ability, making such a comparison unfair (Popham). Yet, even the large achievement test companies have stated that student test scores should not be used to evaluate teachers (Popham). Bracey (2009) noted that many large scale assessments such as NAEP are not aligned with any specific curriculum and provide little or no useful information about individual schools or teacher performance. Many of these large assessments are instructionally insensitive to state or local curriculum (Bracey; Popham, 2007). Another common evaluation scheme is to average all the students’ final grades in class to get a composite score. That composite score is then averaged with all the different classes a teacher has taught to result in a final score that is used to evaluate the teacher’s performance. Cronbach (1970) warned that:
To agglomerate many types of post course performance into a single score is a mistake because failure to achieve one objective is masked by success in another direction. Moreover, since a composite score embodies (and usually conceals) judgments about the importance of various outcomes, only a report that treats the outcomes separately can be useful to educators who have different value hierarchies. (p. 675)

Thus, any attempt to produce a single composite score from student test data may actually conceal student performance. This would provide a false image of a teachers’ effectiveness. Twenty-two years later, Koretz (1992), reiterated a similar point of view: “Simple aggregate scores are not sufficient basis for evaluating education unless they provide enough information to rule out non-educational influences on performance. Most test score databases do not offer that kind of information” (p. 10).

The misuse of test data to evaluate teachers has lead to several problems: cheating and constriction of the curriculum. Using student test scores to evaluate teacher performance has lead teachers into score boosting games they cannot win (Popham, 2000). Cheating on large scale assessment tests by teachers appears to be widespread (Chester, 2005; Smyth, 2008). Chester asserted, “The ethics and professionalism of educators are declining and that decline is due to NCLB” (p. 3). Immense pressure has been placed on teachers to raise test scores at all cost. Many teachers are starting to use test administration practices and preparation that are not appropriate (Rossi, 2002). Popham (2003) and Smyth (2008) noted that teachers and administrators are
caught weekly cheating on large scale assessments. He noted two common
methods of cheating were: teachers pointing out incorrect answers to be re-
examined and allow students additional time to complete tests. As stated earlier,
test preparation and administration practices can increase systemic errors in
assessment data. The increase in error reduces the reliability and validity of the
measurement (Smyth). Perhaps even more alarming is the fact that teachers are
modeling inappropriate behavior for their students (Popham). Boosting test
scores has lead to a mindset of the end justifying the means. This may possibly
lead to the decline of professional ethics in education.

Using student test scores to evaluate schools

Another misuse of assessment has been to use standardized tests to
evaluate schools. It is important to remember that such tests do not have
established valid or reliable scores for the purpose of evaluating schools.
Standardized achievement tests have been designed to provide a meaningful
distribution that allows a comparison of differences among students. Several
studies have shown that as much as 50% of the content of standardized tests is
not taught by most school districts (Popham, 2000). Questions on these tests
tend to be directed toward students of high socioeconomic backgrounds or
students with about average IQ’s (Popham, 2001; Smyth, 2008). The data
produced by these tests reflect socioeconomic status and IQ of students
attending a school rather than how well the school is performing. Steffy and
English (1997) conducted an extensive study of test scores from the National
Assessment of Educational Progress:
Eighty-nine percent of the variance of the scores was explained by four variables: the number of parents living in the home, the parents’ education, community type, and state poverty rate. Tests reflect wealth disparity as opposed to the actual taught school curriculum. Researchers have found that no school-related variables explained variance of test scores that were statistically significant on unaligned tests (p. 6).

This supports the finding of the Coleman Report, a study conducted decades earlier, which found that the socioeconomic status and education level of parents was the greatest predictor of a student’s achievement. Thus, standardized tests are closely linked to the family income (Hursh, 2005; Smyth, 2008). Many studies have clearly shown that a student’s test performance is strongly linked to his/her home life. Since test results are greatly influenced by factors outside the school’s control, it may not be fair to use these scores to evaluate the performance of a school. This increased use of assessment to evaluate schools is a direct result of the accountability movement.

However, research has also documented that lower income students can score as high as upper income students on achievement tests under the right conditions (Edmonds, 1982). Edmonds’ effective school research demonstrated that lower income students could score high on achievement tests if the right school climate could be achieved. The effective school research noted that high achieving lower income schools had similar characteristics. The characteristics of an effective school were: a safe and orderly environment, a clear school mission, instructional leadership, high expectations of student achievement, high
time on tasks, frequent monitoring of student progress, and good community relations (Edmonds).

Schools had to respond to the fact that test scores are being used to evaluate their performance. This misuse of assessment has lead to several problems. First, schools may simply retain more students to keep them from taking high-stake tests. Several states have attempted to do this (Hurst, 2005). Haney (2000) found that the number of ninth grade students in Texas that were progressing to twelfth grade, on-time, decreased dramatically after NCLB. He also noted an increase in the dropout rate of students. Haney stated, “The Texas miracle was really the Texas mirage” (p. 616). He also documented that several states were retaining students to increase test scores.

Teachers, under pressure to raise test scores, begin to use skill and drill test preparation techniques which reduces instructional time for other activities. Thus, teachers do not have time to teach skills and knowledge that are not tested by these tests (Popham, 2003). Many of the skills that teachers once to considered important are no longer being taught (Popham). As test scores become more important, schools actually have reduced what is being taught to students (Hurst, 2005). For example, in Texas, science was dropped from the curriculum in the early grades because it was not tested (Hurst).

Popham (2003) reported teachers have stopped providing enrichment activities, replacing them with skill and drill test preparation activities (Popham, 2003). Many teachers seem to have forgotten that learning is supposed to be fun (Smyth, 2008). Students that have positive learning experiences in school
often become lifelong learners (Smyth). The main goal of the public educational system should be the instruction of children (Popham, 1999). Many of the assessments used by our public school systems are not aligned with the instructional needs of students (Popham, 1999). Popham (2003) contended that the joy of learning has been killed by using the wrong tests to evaluate teachers and schools.

*Schools Manipulate Pass-Fail Scores to Increase Student Achievement*

Another misuse of assessments by schools and school districts is to manipulate cut-scores to allow more students to pass. Hursh (2005) documented that the New York State Department of Education had lowered the cut-scores of many tests in order to allow more students to pass. Several tests that were given only required students to answer 39% of the questions correctly to pass. A similar situation was found in Alabama where students only had to answer approximately 50% of the questions correctly in order to pass the five sections of the Alabama High School Graduation Exam (ADE). Cut-scores are characteristic of criterion-referenced tests. Most experts recommend a cut-score of 75% to 90%. Evidently, the cut scores of several states are well below this recommendation. This gives the teacher, administrators, and public a false image of how well schools are performing.

Merchant and Paulson (2005) analyzed NAEP data and found that 18 states out of 25 with high school graduation exams had lower graduation rates than states that did not have graduation exams, even when controlling for race and socioeconomic status. The study also found negative relationships between
graduation tests and a student’s performance on the SAT. Carnoy (2005) reported that the number of 17 year olds graduating on-time between 1970 and 2000 fell from 76% to 71%. Additionally, Shriberg and Shriberg (2006) reported the graduation rate nationwide was 68% for the class of 2001. The graduation rates appeared to be worse in the South. States in the deep South have some of the lowest graduation rates in the nation (Shriberg & Shriberg). The misuse of assessment may cause a decline in graduation rate across the nation. This misuse could be lowering our students’ preparation for college (Merchant & Paulson).

The negative consequences of high stakes testing such as an increase in the dropout rate, teaching to the test, cultural bias against minorities, and constriction of the curriculum must also be considered. High school graduation rates dropped between 1998 and 2001 in five states studied: 4% in Indiana, 5% in Florida, 3% in New York, 2% in North Carolina, and 1% in South Carolina (Perkins-Gough, 2005). According to Amrein and Berlinear (2002), more than half of the 25 states, after implementing high school graduation exams had an increase in dropouts and GED enrollment.

Studies show that ethnic minorities scored lower than Caucasians on standardized tests with Asian Americans being the exception (Altshuler & Schmartz, 2006). A study by Onwuegbuzie & Daley, 2001, found that on the Standard Assessment of Intelligence (SAI), Caucasians scored 15 points higher than African-Americans and 22 points higher than Hispanics. A study conducted by the College Board on the Scholastic Aptitude Test (SAT) found that
Caucasian students scored an average of 1054, compared to 913 for Hispanic students and 900 for African American students. Some propose this difference may be more of an effect of socioeconomic status rather than race. However, studies have not supported this since there can still be a 12% to 15% gap in test scores when controlling for socioeconomic status (Gandara & Lopez, 1998).

According to Nichols (2003), more than 50% of ethnic minority students in Indiana failed to meet minimum competency standards on the state graduation test for English/language arts. The mathematics test results were even worse with 70% to 80% of minority students failing the test (Nichols). The combined average for English and Math on the Indiana Graduation Exam ranged from 25% to 63% failure for minority students for each graduating class (Nichols). Olson (2006) cautions that as states move to make more alternatives for students who are struggling, it is important that schools do not water-down the curriculum.

The increased use of assessment has resulted in legal challenges of graduation exams. The Standards for Educational and Psychological Testing Standard 13.7 reads, “A decision that will have major impact should not be based on a single test score” (Koretz, 2006, p. 47). Standard 13.6 reads, “When scores are used in determining promotion or graduation, students are to be given a reasonable number of opportunities to pass the test” (Koretz, p. 47). Lawsuits challenging graduation exams based on the above standards were filed in several states including Texas, Florida, and Massachusetts. In Florida, the state’s graduation exam was challenged on the grounds that it violated the constitutional rights of students with disabilities because of inadequate
accommodation while taking the test (Simpson, 2003). The *Debra P. v. Turlington* case challenged the practice of using graduation tests as a basis for awarding diplomas in Florida. The state of Florida proved the content validity of their graduation exam and that the curriculum served as framework to ensure the content was presented to the students (Quilter, 1999). Students were also given multiple chances to pass the test. The court ruled that Florida’s graduation exam was constitutional.

In Texas, a class action suit, *GI Forum et al. v. Texas Education Agency et al.*, was filed on the basis that African Americans and Hispanics failed the graduation exam at a higher rate than Caucasian students (Branch, 2000). The basic issue put before the court was that since African Americans and Hispanics failed at a higher rate, the test discriminated against minorities (Zehr, 2000). Thus, it was proposed that the state’s requirement of passing the graduation exam violated the constitutional rights of minority students (Zehr). “The Texas system was conjunctive: students had to pass a cut-score on the state test, but they also had to meet other requirements for grades in certain courses” (Koretz, 2006, p. 48). As in Florida, students were given multiple chances to pass the test. The United States Federal District Court upheld the state’s right to require a graduation exam and ruled the test did not violate the constitutional rights of minorities (Branch). The court ruled that minority students were not in significant numbers to offset the positive effectiveness of the test (Zehr). The judge also pointed out that the gap in test scores between Caucasians and minorities was closing rapidly (Zehr). Thus, the graduation exam was a major factor in helping
to improve the education quality for minority students (Zehr). The Massachusetts graduation exam was also challenged on the grounds that it discriminated against Hispanics, African Americans, and students with disabilities (Simpson, 2003). The court also upheld the

There have been several other problems resulting from the increase in accountability and high stakes testing: narrowing the curriculum and lack of learning for enrichment or personal development. According to Jacobsen and Rothstein (2006), some teachers must spend 90% of their time teaching just math and reading because these subject areas are tested. This practice allows little time for teaching other subject areas (Jacobsen & Rothstein; Smyth, 2008). Teachers are under pressure to produce results. Many school systems are forcing teachers to prove that everything being taught is linked to testable objectives (Jacobsen & Rothstein). Jacobsen and Rothstein found many teachers reported they were no longer allowed to do anything fun with the students. Many teachers feel that they can no longer take an innovative approach to teaching their students (Smyth). Teacher decision making power has been reduced by the construction of the curriculum leading to a de-professionalization of teaching (Smyth). Some teachers reported this narrowing of the curriculum as a major factor in their leaving the teaching profession (Jacobsen & Rothstein).

The narrowing of the curriculum may disproportionately affect poor students in comparison to wealthy students. According to Jennings and Rentner (2005) in a report for the Center on Education Policy, 97% of high poverty school districts
mandated time requirements for reading. Only half of wealthy school districts mandated minimum time for reading. The report also found that minimum time for electives such as music, art and physical education were reduced. A major impact of No Child Left Behind has been a reduction in the academic balance in the nation’s education system. More experienced teachers echo this feeling with statements such as, “I remember when teaching was fun.”

Another unintended consequence of accountability efforts is that a higher percentage of low-achieving students have been retained or suspended from school at test-taking times to increase test scores. Several states have used achievement test scores instead of course grades as criteria to retain students. For example, in 2006, 24,000 third graders were retained in Florida, and 12,000 third grade students were retained in Texas based on test scores (Allington & McGill-McFranzen, 2006). Many urban school districts have retained high numbers of students as well. Collectively, New York and Miami school districts retained 18,000 third graders who are third time repeaters (Allington & McGill-McFranzen). Retaining 10% of a state’s lowest-achieving students can raise test scores for the next grade level (Allington & McGill-McFranzen).

A study of Florida schools led researchers to believe the same school districts was suspending low achieving students at a higher rate to raise test scores. According to the Figlio’s report, lower achieving students were suspended at a higher rate and for longer periods of time than higher achieving students (Greifner, 2006). The Figlio study involved 504 high, middle, and elementary schools and 41,803 suspension incidents (Greifner). The schools
studied suspended lower-achievers for longer periods of time, 2.35 days, particularly during testing time. Higher-achieving students were suspended for an average of 1.91 days for similar offenses (Greifner).

Summary

The passage of NCLB has led to the development of hundreds of new tests across the United States. Some of these tests have been hastily constructed in order to comply with the law. Many of these new tests are criterion-referenced in nature. CRTs may be more closely aligned with the state and local curriculums than NRTs. These tests, if properly aligned, can provide valuable insight to the successful transmission of curriculum to students. However, teachers and administrators are less familiar with CRTs than NRTs (Frisbie, 2005; Wise et al., 1991). In fact, most teachers and administrators have little or no training in measurement, assessment, and statistics (Popham, 2006a). It is important to remember that the pass/fail score on a criterion-referenced test is called a cut-score. The cut score is set by the individuals developing the test. Thus, unlike the mean of norm-referenced test, which is an actual algebraic property of a set of scores, the cut-score is arbitrarily set. This means that the pass/fail rate and other major indexes, such as standard error and reliability, can be easily manipulated.

Since NCLB requires 100% proficiency of students by 2014, this law has put tremendous pressures on educators to raise test scores. The pressure to raise test scores has led to major misuses of assessments. Teachers and schools are openly cheating to raise scores (Chester, 2005; Smyth, 2008). Test
scores are being used to evaluate teachers and schools even though they were not designed for that purpose (Popham, 2003). Students are being retained and suspended to artificially raise test scores. The curriculum is being narrowed to allow more time for skill and drill to raise test scores (Popham, 2003; Rossi, 2002; Smyth, 2008). Narrowing the curriculum disproportionately affects poor schools and students (Jacobsen & Rothstein, 2006).

The lack of assessment literacy has led to the misuse of assessment in a way that damages teachers, students, administrators, and schools. However, improving assessment literacy of teachers and administrators has the potential to provide dramatic increase in learning that provides a real and positive experience for students.
CHAPTER III

METHODOLOGY

Summary of Study

This study investigated the assessment literacy of teachers and administrators in regard to criterion-referenced tests. For the purpose of this study, administrators included principals and assistant principals within a school. First, the study determined the degree of training in measurement, assessment, and statistics that teachers and administrators have received. Second, the study investigated the accuracy in which teachers and administrators identified the theoretical differences of norm-referenced and criterion-referenced tests. Third, the study determined the accuracy of teachers and administrators in identifying the major concepts of validity and reliability. Fourth, the study determined the ability of teachers and administrators to recognize common misuse of test data. The basic research design of this study was correlational.

Research Questions

1. How much training do teachers and administrators report having in measurement, assessment, and statistics?

2. Can teachers and administrators identify the theoretical differences between norm-referenced and criterion-referenced tests?

3. Can teachers and administrators identify the major concepts of reliability and validity?
4. Does method of certification or highest degree held make a difference in the training of school personnel in measurement, assessment, and statistical reasoning?

5. Can teachers and administrators identify potential misuses of assessment?

6. What are teacher’s attitudes on the application of educational statistics and their use in education?

7. Will there be a difference in assessment literacy on norm-referenced/criterion referenced tests, validity/reliability, and potential misuses between teachers and administrators?

**Participants**

Stratified sampling was used to select the participants from the population of teachers and administrators in the state of Alabama and Mississippi. The sample was stratified into two main areas. The first area contained the coastal counties of Alabama. The coastal counties of Alabama included Mobile and Baldwin Counties. The participants from Alabama were selected from Mobile, Baldwin, and Saraland School Districts. The coastal area of Mississippi was made up of Jackson, Harrison, and Hancock counties. The participants from Jackson County were selected from Moss Point Separate, Ocean Springs, and Pascagoula Schools Districts. The participants from Harrison County were selected from Biloxi Public, Gulfport Public, Long Beach Public, and Pass Christian Public School Districts. The participants from Hancock County were selected from Bay St. Louis and Waveland School Districts.
The participants in this study were teachers and administrators currently practicing in the states of Alabama and Mississippi. A G-Power analysis was used to determine sample size for the study. The ANOVA G-Power analysis was run with an effect size of 0.25, alpha 0.05, power 0.95, and 4 groups. A sample size of 280 participants was calculated by the program.

A second G-Power analysis was run for a MANOVA. The same effect size, alpha, power, and number of groups were used, and a sample size of 48 was calculated by the program. The sample size of 280 was used during the study because it allowed for a more complex analysis of the data, as well as of the instrument itself.

**Instrument**

The Criterion-Referenced Assessment Questionnaire (Appendix A) to be used during this study was developed by the researcher. The instrument contained 4 main sections. The first section solicited basic demographic information such as: certification, degree, gender, grade level, title, years experience and coursework. The section included 10 select response items. The second section measured theoretical constructs of the study and included 24 multiple choice questions. Nine questions: 1, 3, 4, 5, 6, 7, 10, 14, and 19, assessed the accuracy of teachers and administrators to identify the theoretical differences between NRTs and CRTs. Eight questions: 2, 8, 9, 11, 12, 15, 23, and 24, assessed the accuracy of teachers and administrators to identify major concepts of reliability and validity. Seven questions: 13, 16, 17, 18, 20, 21, and 22, assessed whether teachers and administrators can identify potential misuses
of assessment. The third section included 10 Likert scale type items meant to measure the perception of teachers and administrators on the use of assessment, measurement, and statistical reasoning in education. The Likert scale for questions 25-34 were as follows: Strongly Disagree (SD=1), Disagree (D=2), Not Sure (NS=3), Agree (A=4), and Strongly Agree (SA=5).

The instrument was reviewed by a panel of experts who helped in the process of developing the instrument and worked to increase the face and content validity of the instrument. The panel consisted of four experts in the area of educational measurement. Two members of the panel were university professors who taught measurement and statistics. The other two members of the panel were education practitioners. One is currently an elementary principal with over 30 years educational experience and a PhD in educational administration. The other practitioner was a high school teacher with a variety of educational experiences, including working for the Alabama State Department of Education, a private educational company, as well as holding a PhD in educational administration. The purpose of this particular panel was to ensure that both practitioners and college professors of measurement and statistics were included. Also, expertise could be provided from both an elementary and secondary school perspective.

Each member of the panel was provided with a copy of the Criterion-Referenced Assessment Questionnaire and was asked to provide feedback to help refine the questions to increase the face and content validity of the instrument. The panel returned the Criterion-Referenced Assessment
Questionnaire with corrections indicated by each member of the panel. The researcher then adjusted the Criterion-Referenced Assessment Questionnaire to address the concerns of the panel.

The Criterion-Referenced Assessment Questionnaire was pilot tested in order to establish reliability and validity of the instrument. The instrument was piloted in a coastal elementary school in Mississippi. This particular school was chosen because of accessibility and location. The principal is a part time college professor who teaches classes in educational research. The sample in the pilot study was appropriate because the school was in the same coastal area where the study occurred and included approximately 33 teachers and administrators, grades K–5. This school was excluded from the sample population used during the study.

Data was gathered and analyzed during the pilot study in order to improve the instrument. Cronbach’s alpha for internal consistency was established during the pilot study. Any items that were left blank or misunderstood by the participant were identified during this process. The researcher and panel of experts worked to refine the instrument to meet established levels for reliability and validity. The instrument was then be prepared for use and distribution in the study.

Procedure

The researcher sought approval to conduct this research from the Institutional Review Board at the University of Southern Mississippi. Once approval was granted (Appendix B), the study commenced. The participants of this study were selected from the coastal counties of the states of Alabama and
Mississippi. The coastal counties of Alabama were Mobile and Baldwin Counties. The coastal counties were determined using the Alabama School District Map (Appendix C). The school districts in coastal Alabama were Mobile and Baldwin County School Districts, as well as Saraland School District. The coastal counties of Mississippi (Appendix D) were Jackson, Harrison, and Hancock Counties. The school districts in Jackson County were Jackson County, Moss Point, Ocean Springs, and Pascagoula Schools. The school districts in Harrison County were Biloxi, Gulfport, Long Beach, and Pass Christian Schools. The school districts in Hancock County were Bay St. Louis and Waveland School Districts. A directory of Alabama and Mississippi schools were downloaded from the Alabama and Mississippi State Department of Education’s websites. The Alabama and Mississippi School Information Directory were used to compile a list of the school districts’ addresses and phone numbers within each coastal region. Each school district’s superintendent was contacted by mail or e-mail (Appendix E) to ask permission to take part in the study. The number of school districts that took part in the study were determined by how many voluntarily agree to participate. Once the school district agreed to take part in the study, individual schools within school districts were contacted by mail or e-mail (Appendix F) to obtain permission to take part in the study. If a school declined to take part in the study, another school was systematically selected from the list generated previously. This process continued until the end of the 2009-2010 school year and a sufficient number (280) of participants responded to the questionnaire in order to reach the sample size, as required by
the G-Power analysis, required for this study. Once a school agreed to take part in the study, an appropriate number of the Criterion-Referenced Assessment Questionnaires and Informed Consent Letters (Appendix G) were provided to all participants and mailed to each school. The questionnaire was administered at a faculty meeting and collected at the end of the meeting. If the school was within driving distance (approximately 50 miles) the researcher administered and collected the questionnaire. If the school was not within this driving distance, the school’s principal or assistant principal was asked to administer and collect the questionnaire collectively. Once completed, the questionnaires were returned to the researcher collectively in a self-addressed, stamped, envelope. Once the questionnaires were received, they were entered into SPSS software for analysis.

Analysis

The instrument was analyzed using Exploratory Factor Analysis. A Principal Components Analysis was run to determine the underlying structure of the Likert items on the Criterion-Referenced Assessment Questionnaire. During this analysis the underlying structure of the Likert items of the questionnaire were established as well as the overall reliability of the instrument. The three main theoretical constructs of the Likert items were educator’s attitudes toward the use of statistics in education, educator training in the area of measurement, and whether the schools use data that is reliable and valid.

In order to address Research Questions 1–7, SPSS software was utilized to analyze the data gathered during this research project. In addressing
Research Question 1, information was collected on the amount of measurement, assessment, and statistic training that the participants have received about the demographic section of the instrument. The data were analyzed using descriptive statistics for frequencies and percentages were reported.

In addressing Research Question 2, information on the ability of teachers and administrators to identify the theoretical differences between NRTs and CRTs was gathered using the Criterion-Referenced Assessment Questionnaire. The data were analyzed and the means and standard deviations will be reported.

In addressing Research Question 3, information on the ability of teachers and administrators to identify basic concepts of reliability and validity was gathered using the Criterion-Referenced Assessment Questionnaire. The data were analyzed and the mean and standard deviation were reported.

In addressing Research Question 3, information on the ability of teachers and administrators to identify basic concepts of reliability and validity was gathered using the Criterion-Referenced Assessment Questionnaire. The data were analyzed and the mean and standard deviation were reported.

In order to address Research Question Number 4, information on the method of certification or highest degree held making a difference in the training of school personnel in measurement, assessment and statistical reasoning, data were gathered using the Criterion-Referenced Assessment Questionnaire. The data were analyzed and the mean and standard deviation were reported.

In order to address Research Question Number 5, information on the teachers’ and administrators’ ability to identify potential misuse of assessment
was gathered using the Criterion-Referenced Assessment Questionnaire. The data were analyzed and mean and standard deviation was reported.

In order to address Research Question Number 6, information on teacher’s attitudes on the application of educational statistics and their uses in education, was gathered using the Criterion-Referenced Assessment Questionnaire. The data were analyzed and mean and standard deviation were reported.

In order to address Research Question Number 7, information on the difference in assessment literacy on NRTs/CRTs, validity/reliability, and potential misuses between teachers and administrators, was gathered using the Criterion-Referenced Assessment Questionnaire. The data were analyzed and mean and standard deviation were reported.

In addressing Research Hypothesis 1, 2, and 3, an analysis of variance (ANOVA) was used to determine if there was a significant difference on performance on the Criterion-Referenced Assessment Questionnaire between teachers and administrators based on years of experience, method of certification, and highest degree held. A significant ANOVA was followed up by a post-hoc test using a Bonferroni Correction.

In regard to Research Hypothesis 1, a significant ANOVA was followed up by a post-hoc test using a Bonferroni Correction to determine if there is a significant difference in assessment literacy based on highest degree held.
In regard to Research Hypothesis 2, a significant ANOVA was followed up by a post-hoc test using a Bonferroni Correction to see if there is a significant difference in assessment literacy based on type of certification.

In regard to Research Hypothesis 3 a significant ANOVA was followed up by a post-hoc test using a Bonferroni Correction to determine if there is a significant difference in assessment literacy based on years experience.
CHAPTER IV

ANALYSIS OF DATA

Simple Stratified sampling was used to select the participants from the population of teachers and administrators in the states of Alabama and Mississippi. Four school districts, two from each state, agreed to take part in the study. Twenty-one schools were surveyed during the study. A total of 1720 questionnaires were distributed during the study. The questionnaire was completed by 380 educators. The participants from Alabama returned 226 questionnaires. The participants from Mississippi returned 154 questionnaires. The overall return rate was 22%.

The sample used for this analysis included 310 female and 70 male educators. Educators’ grade level taught ranged kindergarten through twelfth grade. Middle school grades six to eight represented the smallest number of participants with 54 educators. Secondary schools grades nine to twelve represented the largest number of participants with 205 educators. Jobs of the participants represented in the sample included 18 administrators, 10 counselors, and 352 teachers. Demographic information of participants are reported on next page (see Table 1).
Table 1

Demographic Information of Participants

<table>
<thead>
<tr>
<th>Demographic information</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>70</td>
<td>18.4</td>
</tr>
<tr>
<td>Female</td>
<td>310</td>
<td>81.6</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Total</td>
<td>380</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Job</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>352</td>
<td>92.6</td>
</tr>
<tr>
<td>Administrator</td>
<td>18</td>
<td>4.7</td>
</tr>
<tr>
<td>Counselor</td>
<td>10</td>
<td>2.6</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Total</td>
<td>380</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Grade</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-5</td>
<td>119</td>
<td>31.2</td>
</tr>
<tr>
<td>6-8</td>
<td>54</td>
<td>14.2</td>
</tr>
<tr>
<td>9-12</td>
<td>205</td>
<td>53.9</td>
</tr>
<tr>
<td>Missing</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>Total</td>
<td>380</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Instrument

The Criterion-Referenced Assessment Questionnaire used for this study was developed by the researcher. The instrument contained three main sections. The first section solicited basic demographic information such as: certification, degree, gender, grade level, title, years experience, and coursework. This section included 10 select response items. The second section measured a participant’s knowledge of assessment and included 24 multiple choice questions. Nine questions: 1, 3, 4, 5, 6, 7, 10, 14, and 19, assessed the accuracy of teachers and administrators in identifying the theoretical differences between NRTs and CRTs. Eight questions: 2, 8, 9, 11, 12, 15, 23, and 24, assessed the accuracy of teachers and administrator’s ability to identify major concepts of reliability and validity. Seven questions: 13, 16, 17, 18, 20, 21, and 22, assessed whether teachers and administrators could identify potential misuses of assessment. The third section included 10 Likert scale items meant to measure the beliefs of teachers and administrators on the use of assessment, measurement, and statistical reasoning in education. The Likert scale for questions 25-34 was as follows: Strongly Disagree (SD=1), Disagree (D=2), Not Sure (NS=3), Agree (A=4), and Strongly Agree (SA=5).

Reliability of Criterion-Referenced Questionnaire

A pilot study was conducted to determine the reliability of the Criterion-Referenced Questionnaire. The sample in the pilot study was 32 (29 females and three males). The alpha coefficient was .403 for questions 1-24, the knowledge based items. The alpha coefficient was .321 on questions 25-34 the Likert scale items. Both alpha coefficients were well below the .70, which is what
is acceptable for educational research. The panel of experts reviewed the reliability of the instrument determined by the pilot study and determined that the reliability would most likely improve with a larger sample size. Since the content and face validity of the instrument seemed sound the panel agreed the study should proceed.

At the completion of data collection, the alpha coefficient was calculated for the Criterion-Referenced Questionnaire. The alpha coefficient was .734 for the knowledge and .777 for Likert items. Both were above .70, which is what is acceptable for educational research.

Exploratory Factor Analysis was conducted on the Likert (beliefs) portion of the instrument. The information provided by this analysis is meant to provide a complete descriptive analysis of the instrument and was not used in subsequent hypothesis testing. A principal component analysis was run on the Criterion-Referenced Questionnaire to determine the structure of the instrument. Item correlations ranged from .620 to -.237. Question 26 had the highest mean of 3.45. Question 27 had the highest standard deviation of 1.20. Question 33 had the lowest mean of 1.95. Question 32 had the lowest standard deviation of 1.06. Mean and standard deviation of items appeared to be within normal range (see Table 2).
Table 2

*Descriptive Statistics Likert Items (N=354)*

<table>
<thead>
<tr>
<th>Item</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Not Sure</th>
<th>Agree</th>
<th>Strongly Agree</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>29. School improvement based on data</td>
<td>(50)</td>
<td>(77)</td>
<td>(119)</td>
<td>(108)</td>
<td>(18)</td>
<td>2.91</td>
<td>1.12</td>
</tr>
<tr>
<td>30. Believe test district use reliable/valid</td>
<td>(58)</td>
<td>(96)</td>
<td>(92)</td>
<td>(101)</td>
<td>(26)</td>
<td>2.85</td>
<td>1.19</td>
</tr>
<tr>
<td>27. Adequate training in statistics</td>
<td>(67)</td>
<td>(91)</td>
<td>(93)</td>
<td>(96)</td>
<td>(24)</td>
<td>2.79</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>17.6%</td>
<td>23.9%</td>
<td>24.4%</td>
<td>25.2%</td>
<td>6.3%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 (continued).

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>28. School improvement can be implemented</td>
<td>(44)</td>
<td>(121)</td>
<td>(103)</td>
<td>(82)</td>
<td>(23)</td>
<td>2.74</td>
</tr>
<tr>
<td>District offers inservice in measurement</td>
<td>(68)</td>
<td>(98)</td>
<td>(124)</td>
<td>(61)</td>
<td>(9)</td>
<td>2.58</td>
</tr>
<tr>
<td>Standard for educational testing</td>
<td>(94)</td>
<td>(82)</td>
<td>(101)</td>
<td>(75)</td>
<td>(9)</td>
<td>2.52</td>
</tr>
<tr>
<td>District show that test are reliable / valid</td>
<td>(86)</td>
<td>(104)</td>
<td>(94)</td>
<td>(62)</td>
<td>(14)</td>
<td>2.49</td>
</tr>
<tr>
<td>Students' test used to evaluate teacher</td>
<td>(167)</td>
<td>(90)</td>
<td>(61)</td>
<td>(36)</td>
<td>(6)</td>
<td>1.95</td>
</tr>
</tbody>
</table>

Note: The size of the column in the table did not allow for the entire Likert items to be included. See Appendix A for a full explanation of each Likert item. The Likert scale for questions 25-34 were as follows: Strongly Disagree (SD=1), Disagree (D=2), Not Sure (NS=3), Agree (A=4), and Strongly Agree (SA=5).
The Kaiser-Meyer-Olkin Measure (KMO) of sampling adequacy was .788. Bartlett’s Test of Sphericity was significant. The KMO and Bartlett’s Test were within acceptable values to continue with the extraction of the factors. Several criteria were used to help determine the number of factors in the Criterion-Referenced Questionnaire. First, an initial extraction was conducted using a direct Oblimin Solution on the survey to determine the number of factors that had Eigenvalues greater than one. Two factors appeared to be present using this rule. Eigenvalues ranged from 1.564 to 3.583. Next, the commonalities were assessed to determine the loading of the items for each factor. Question twenty-eight had the lowest loading with .476. The question was retained in the analysis.

A Scree Plot was used to help determine the number of factors present in the Criterion-Referenced Questionnaire. A visual inspection in the Scree Plot showed four possible factors. An extraction was run forcing the computer to use 4 factors. The total variability explained by the model increased from 51.469 to 69.529%. However, only one item loaded on the fourth factor. So, the fourth factor could not be used in the final extraction.

Ultimately, the extraction yielded three factors. Factor one included questions 29, 30, 31, and 33 and dealt with school use of data that is reliable and valid. Factor two included questions 25, 26, and 28 and dealt with attitudes toward the use statistics in education. Factor three included questions 27, 32, 34, and dealt with educator training in measurement (see Table 3).
### Table 3

*Pattern Matrix Criterion-Referenced Questionnaire*

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Component</th>
</tr>
</thead>
</table>
| 33. students’ test used to evaluate teachers                                   | 1   
| 29. school improvement based on data                                           | 2   
| 30. believe test district use reliable / valid                                  | 3   
| 31. believe test district use reliable / valid                                  | 1   
| 26. statistical analysis of data                                               | 29.0   
| 25. opinion educational statistics                                             | 29.0   
| 28. school improvement can be implemented without analysis data                | 29.0   
| 27. adequate training in statistics                                            | 29.0   
| 32. district offers inservice in measurement                                    | 29.0   
| 34. standards for educational testing                                          | 29.0   

Total variability explained by the factors was 61.153%. Factor 1 explained the largest amount of variance with 35.828%. Factor 2 was 15.641%. Factor 3 was 9.684%. The rotated solution yielded comparable results to the unrotated solution. The data was reanalyzed using the three factors with Varimax rotation.
The values of the structure and pattern obtained decreased over the direct Oblimin Solutions. The Direct Oblimin Solution had better structure and pattern compared to Varimax. The amount of variability explained by the Direct Oblimin Rotation was 61.153%. The factor analysis portion of the study was conducted for no purpose other than initiating the development of a questionnaire that may have utility in subsequent studies.

Factor one had an alpha coefficient of .733, M = 2.55, and SD = 3.38. Factor two had an alpha coefficient of .081, M = 3.05, and SD = 1.94. Factor three had an alpha coefficient of .668, M = 2.62, and SD = 2.64.

Results

In order to determine if there was a difference in assessment literacy on norm-referenced/criterion-referenced tests, validity/reliability, and potential misuses between teachers and administrators, means of the groups were examined. Teachers scored the lowest on the Criterion-Referenced Questionnaire M=57.022 SD=18.743. Counselors score the highest on the Criterion-Referenced Questionnaire M=66.166 SD=16.445. The average score on the questionnaire was M=57.415 SD=18.985. There did not appear to be a significant difference between the three groups (see Table 4).
Table 4

*Descriptive Statistics for Job Type (N=380)*

<table>
<thead>
<tr>
<th>Job Type</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>57.02</td>
<td>18.74</td>
<td>352</td>
</tr>
<tr>
<td>Counselor</td>
<td>66.30</td>
<td>16.44</td>
<td>10</td>
</tr>
<tr>
<td>Administrator</td>
<td>60.16</td>
<td>23.22</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>57.41</td>
<td>18.93</td>
<td></td>
</tr>
</tbody>
</table>

In order to determine if teachers and administrators could identify the theoretical differences between NRTs and CRTs, descriptive information was collected and analyzed. Questions on the knowledge section of the Criterion-Referenced Questionnaire that dealt with theoretical differences were questions: 1, 3, 4, 5, 6, 7, 10, 14, and 19. Teachers and administrators correctly answered 47.36% of the questions in this subsection of the questionnaire. Question 7 was answered with the highest accuracy with 64.6% of the respondents answering correctly. Question 19 had the least number of correct responses with 23.1% correctly answering the question (see Table 5).

In order to determine if teachers and administrators could identify the major concepts of reliability and validity, descriptive information was gathered and analyzed. Questions on the knowledge section of the Criterion-Referenced Questionnaire that dealt with reliability and validity were: 2, 8, 9, 11, 12, 15, 23, and 24. Teachers and administrators correctly answered 58.68% of the
questions in this subsection of the questionnaire. Question 12 was answered with the highest accuracy with 89.2% of the respondents answering correctly. Question 2 had the least number of correct responses with 18.1% correctly answering the question (see Table 5).

In order to determine if teachers and administrators could identify potential misuses of assessment, descriptive information was gathered and analyzed. Questions on the knowledge section of the Criterion-Referenced Questionnaire that dealt with potential misuses were: 13, 16, 17, 18, 20, 21, and 22. Teachers and administrators correctly answered 66.68% of the questions in this subsection of the questionnaire. Question 13 was answered with the highest accuracy with 84.8% of the respondents answering correctly. Question 16 had the least number of correct responses with 47.8% correctly answering the question (see Table 5).

Table 5

*Descriptive Statistic Criterion-Referenced Questionnaire (N=380)*

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Correct</th>
<th>Incorrect</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.</td>
<td>(246) 64.6%</td>
<td>(130) 34.1%</td>
<td>376</td>
</tr>
<tr>
<td>14.</td>
<td>(241) 63.3%</td>
<td>(135) 35.4%</td>
<td>376</td>
</tr>
<tr>
<td>10.</td>
<td>(222) 58.3%</td>
<td>(153) 40.2%</td>
<td>375</td>
</tr>
<tr>
<td>6.</td>
<td>(205) 53.8%</td>
<td>(169) 44.4%</td>
<td>374</td>
</tr>
<tr>
<td>1.</td>
<td>(205) 53.8%</td>
<td>(172) 45.1%</td>
<td>377</td>
</tr>
</tbody>
</table>
Table 5 (continued).

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5. (170)</td>
<td>44.6%</td>
<td>(206)</td>
<td>54.1%</td>
<td>376</td>
<td></td>
</tr>
<tr>
<td>3. (134)</td>
<td>35.5%</td>
<td>(243)</td>
<td>64.5%</td>
<td>377</td>
<td></td>
</tr>
<tr>
<td>4. (114)</td>
<td>29.9%</td>
<td>(262)</td>
<td>68.8%</td>
<td>376</td>
<td></td>
</tr>
<tr>
<td>19. (88)</td>
<td>23.1%</td>
<td>(287)</td>
<td>75.3%</td>
<td>375</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>47.3%</td>
<td>52.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reliability and validity questions

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>12. (340)</td>
<td>89.2%</td>
<td>(36)</td>
<td>9.4%</td>
<td>376</td>
<td></td>
</tr>
<tr>
<td>15. (327)</td>
<td>85.8%</td>
<td>(49)</td>
<td>12.9%</td>
<td>376</td>
<td></td>
</tr>
<tr>
<td>23. (295)</td>
<td>77.4%</td>
<td>(81)</td>
<td>21.3%</td>
<td>376</td>
<td></td>
</tr>
<tr>
<td>24. (294)</td>
<td>77.2%</td>
<td>(80)</td>
<td>21.0%</td>
<td>374</td>
<td></td>
</tr>
<tr>
<td>11. (198)</td>
<td>52.0%</td>
<td>(178)</td>
<td>46.7%</td>
<td>376</td>
<td></td>
</tr>
<tr>
<td>8. (165)</td>
<td>43.3%</td>
<td>(210)</td>
<td>55.1%</td>
<td>375</td>
<td></td>
</tr>
<tr>
<td>9. (100)</td>
<td>26.2%</td>
<td>(275)</td>
<td>72.5%</td>
<td>375</td>
<td></td>
</tr>
<tr>
<td>2. (70)</td>
<td>18.1%</td>
<td>(307)</td>
<td>80.6%</td>
<td>377</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>58.6%</td>
<td>41.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Potential misuses questions

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>13. (323)</td>
<td>84.8%</td>
<td>(53)</td>
<td>13.9%</td>
<td>376</td>
<td></td>
</tr>
<tr>
<td>17. (317)</td>
<td>83.2%</td>
<td>(60)</td>
<td>15.7%</td>
<td>377</td>
<td></td>
</tr>
<tr>
<td>18. (270)</td>
<td>70.9%</td>
<td>(105)</td>
<td>27.6%</td>
<td>375</td>
<td></td>
</tr>
<tr>
<td>20. (240)</td>
<td>63.0%</td>
<td>(135)</td>
<td>35.4%</td>
<td>375</td>
<td></td>
</tr>
<tr>
<td>21. (237)</td>
<td>62.2%</td>
<td>(139)</td>
<td>36.5%</td>
<td>376</td>
<td></td>
</tr>
<tr>
<td>22. (209)</td>
<td>54.9%</td>
<td>(166)</td>
<td>43.6%</td>
<td>375</td>
<td></td>
</tr>
</tbody>
</table>
Coursework in Measurement, Statistics, and Research Design by Job

In order to determine how much training teachers and administrators had in measurement, assessment, and statistics, descriptive information was collected and analyzed. Administrators reported the highest percentage of coursework taken: 27.7% = 1 class, 33.3% = 2-3 classes, 22.2% = 4-5 classes and 11.1% = 6 or more classes. Counselors reported the lowest percentage of coursework in Educational Measurement: 40% = 1 class, 30% = 2-3 classes, 0% = 4-5 classes and 0% = 6 or more classes. Frequencies and percentages for courses were reported in Table 6.

Table 6

Descriptive Statistics Measurement Courses (N=378)

<table>
<thead>
<tr>
<th>Job Type</th>
<th>0</th>
<th>1</th>
<th>2-3</th>
<th>4-5</th>
<th>6+</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>(46)</td>
<td>(114)</td>
<td>(145)</td>
<td>(33)</td>
<td>(12)</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>13.1%</td>
<td>32.6%</td>
<td>41.4%</td>
<td>9.4%</td>
<td>3.4%</td>
<td></td>
</tr>
<tr>
<td>Counselor</td>
<td>(3)</td>
<td>(4)</td>
<td>(3)</td>
<td>(0)</td>
<td>(0)</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>30.0%</td>
<td>40.0%</td>
<td>30.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td></td>
</tr>
<tr>
<td>Administrators</td>
<td>(1)</td>
<td>(5)</td>
<td>(6)</td>
<td>(4)</td>
<td>(2)</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>5.5%</td>
<td>27.7%</td>
<td>33.3%</td>
<td>22.2%</td>
<td>11.1%</td>
<td></td>
</tr>
</tbody>
</table>
The number of Educational Statistics classes taken was analyzed next.

Teachers reported the highest percentage of Educational Statistics classes taken: 38.9%=1 class, 29.3%=2-3 classes, 1.8%=4-5 classes and 1.1%=6 or more classes. Counselors reported the lowest percentage of Educational Statistics classes taken: 40%=0 classes, 20%=1 class, 40%=2-3 classes, 0%=4-5 classes and 0%=6 or more classes. Frequencies and percentages for courses for courses in Educational Statistics were reported in Table 7.

Table 7

*Descriptive Educational Statistics Courses (N=379)*

<table>
<thead>
<tr>
<th>Educational Statistics Courses</th>
<th>0</th>
<th>1</th>
<th>2-3</th>
<th>4-5</th>
<th>6+</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Job Type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>(101)</td>
<td>(137)</td>
<td>(103)</td>
<td>(7)</td>
<td>(4)</td>
<td>352</td>
</tr>
<tr>
<td></td>
<td>28.6%</td>
<td>38.9%</td>
<td>29.3%</td>
<td>1.98%</td>
<td>1.1%</td>
<td></td>
</tr>
<tr>
<td>Counselor</td>
<td>(40)</td>
<td>(2)</td>
<td>(4)</td>
<td>(0)</td>
<td>(0)</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>40.0%</td>
<td>20.0%</td>
<td>40.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td></td>
</tr>
<tr>
<td>Administrators</td>
<td>(5)</td>
<td>(2)</td>
<td>(7)</td>
<td>(0)</td>
<td>(3)</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>29.4%</td>
<td>11.7%</td>
<td>41.2%</td>
<td>0.0%</td>
<td>17.6%</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>(110)</td>
<td>(141)</td>
<td>(114)</td>
<td>(7)</td>
<td>(7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29.0%</td>
<td>37.2%</td>
<td>30.0%</td>
<td>1.8%</td>
<td>1.8%</td>
<td></td>
</tr>
</tbody>
</table>
The number of Math Statistics courses taken was analyzed next. Administrators reported the highest average percentage of coursework taken: 22.2\%=0 classes, 33.3\%=1 class, 27.7\%=2-3 classes, 5.5\%=4-5 classes and 11.1\%=6 or more classes. Counselors reported the lowest number of coursework taken. Frequencies and percentages for courses in Math Statistics were reported in Table 8.

**Table 8**

*Descriptive Math Statistics Course by Job (N=380)*

<table>
<thead>
<tr>
<th>Job Type</th>
<th>Math Statistics Courses</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Teacher</td>
<td>36.6%</td>
<td>32.1%</td>
</tr>
<tr>
<td>Counselor</td>
<td>50.0%</td>
<td>30.0%</td>
</tr>
<tr>
<td>Administrators</td>
<td>22.2%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Total</td>
<td>36.3%</td>
<td>32.1%</td>
</tr>
</tbody>
</table>

The number of Research Design courses taken was analyzed next. Administrators reported the following percentage of coursework taken: 50\%=0 classes, 5.55\%=1 class, 22.22\%=2-3 classes, 11.11\%=4-5 classes and 11.11\%=6 or more. Teachers reported the following percentage of coursework taken: 49.85\%=0 classes, 28.2\%=1 class, 17.66\%=2-3 classes, 3.7\%=4-5
classes and 0.56%=6 or more. Frequencies and percentages were reported in Table 9.

Table 9

Descriptive Statistics Research Design Courses (N=379)

<table>
<thead>
<tr>
<th>Job Type</th>
<th>0</th>
<th>1</th>
<th>2-3</th>
<th>4-5</th>
<th>6+</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>175</td>
<td>99</td>
<td>62</td>
<td>13</td>
<td>2</td>
<td>351</td>
</tr>
<tr>
<td>Counselor</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Administrators</td>
<td>9</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>189</td>
<td>102</td>
<td>68</td>
<td>15</td>
<td>5</td>
<td>379</td>
</tr>
</tbody>
</table>

Coursework in Measurement, Statistics, and Research Design by Certification

In order to determine if the type of certification influenced the amount of training an educator received in measurement, assessment and statistics, descriptive information was collected and analyzed. Traditionally, certified teachers reported a slightly higher percentage of measurement coursework taken: 13.0%=0 classes, 28.96%=1 class, 42.6%=2-3 classes, 11.37%=4-5 classes and 4.48%=6 or more classes. Alternative certified teachers reported the lowest percentage of coursework taken: 13.63%=0 classes, 44.1%=1 class,
36.36% = 2-3 classes, 4.54% = 4-5 classes and 1.14% = 6 or more classes.

Frequencies and percentages are reported in Table 10.

Table 10

Descriptive Statistics Measurement Courses Certification (N=378)

<table>
<thead>
<tr>
<th>Certification</th>
<th>Measurement courses</th>
<th>Traditional</th>
<th>Alternative</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>(38) 13.10%</td>
<td>(12) 13.63%</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>(84) 28.96%</td>
<td>(39) 44.31%</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>2-3</td>
<td>(122) 42.06%</td>
<td>(32) 36.36%</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>4-5</td>
<td>(33) 11.37%</td>
<td>(4) 4.54%</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>6+</td>
<td>(13) 4.48%</td>
<td>(1) 1.14%</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>(290) 100.00%</td>
<td>(88) 100.00%</td>
<td></td>
</tr>
</tbody>
</table>

Traditionally certified teachers reported taking a higher number of Educational Statistic classes: 27.49% = 0 classes, 38.14% = 1 class, 29.8% = 2-3 classes, 2.40% = 4-5 classes and 2.06% = 6 or more classes. Alternatively certified teachers reported: 34.09% = 0 classes, 34.09% = 1 class, 30.68% = 2-3 classes, 0% = 4-5 classes and 1.3% = 6 or more classes. Frequencies and percentages are reported in Table 11.
Alternatively certified teachers reported taking a higher number of math statistic courses: 30.68%=0 classes, 32.54%=1 class, 25.00%=2-3 classes, 9.0%=4-5 classes and 2.27%=6 or more. Traditionally certified teachers reported taking a lower number of classes: 38.01%=0 classes, 31.84%=1 class, 26.02%=2-3 classes, 2.05%=4-5 classes 2.05%=6 or more classes. Frequencies and percentages are reported in Table 12.
Table 12

*Descriptive Math Statistics Courses Certification (N=380)*

<table>
<thead>
<tr>
<th>Math Statistics courses</th>
<th>Traditional</th>
<th>Alternative</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(111) 38.01%</td>
<td>(27) 30.68%</td>
<td>138</td>
</tr>
<tr>
<td>1</td>
<td>(93) 31.84%</td>
<td>(29) 32.54%</td>
<td>122</td>
</tr>
<tr>
<td>2-3</td>
<td>(76) 26.02%</td>
<td>(22) 25.00%</td>
<td>98</td>
</tr>
<tr>
<td>4-5</td>
<td>(6) 2.05%</td>
<td>(8) 9.09%</td>
<td>14</td>
</tr>
<tr>
<td>6+</td>
<td>(6) 2.05%</td>
<td>(2) 2.27%</td>
<td>8</td>
</tr>
<tr>
<td>(292) 100.00%</td>
<td>(88) 100.00%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alternatively certified teachers reported a slightly higher number of Research Design coursework taken: 46.59%=0 classes, 23.86%=1 class, 23.86%=2-3 classes, 4.54%=4-5 classes and 1.13%=6 or more classes. Traditionally certified teachers reported taking: 50.85%=0 classes, 27.73%=1 class, 16.15%=2-3 classes, 3.78%=4-5 classes and 1.37%=6 or more classes. Frequencies and percentages are reported in Table 13.
Table 13

Descriptive Statistics Research Design Courses Certification (N=379)

<table>
<thead>
<tr>
<th>Certification</th>
<th>Research Design courses</th>
<th>Traditional</th>
<th>Alternative</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(148) 50.85%</td>
<td>(41) 46.59%</td>
<td></td>
<td>189</td>
</tr>
<tr>
<td>1</td>
<td>(81) 27.73%</td>
<td>(21) 23.86%</td>
<td></td>
<td>102</td>
</tr>
<tr>
<td>2-3</td>
<td>(47) 16.15%</td>
<td>(21) 23.86%</td>
<td></td>
<td>68</td>
</tr>
<tr>
<td>4-5</td>
<td>(11) 3.78%</td>
<td>(4) 4.54%</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>6+</td>
<td>(4) 1.37%</td>
<td>(1) 1.13%</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>(291) 100.00%</td>
<td>(88) 100.00%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coursework in Measurement, Statistics, and Research Design by Degree

In order to determine if the highest degree held influenced the amount of training an educator received in measurement, assessment, and statistics, descriptive information was collected and analyzed. Educators with a bachelor’s degree reported the lowest number of coursework taken: 17.1%=0 classes, 36.89%=1 class, 40.41%=2-3 classes, 2.05%=4-5 classes and 3.425%=6 or more classes. Educators with a doctorate degree reported the highest number of measurement coursework taken: 0%=0 classes, 0%=1 class, 40%=2-3 classes, 20%=4-5 classes and 40%=6 or more classes. Frequencies and percentages are reported in Table 14.
Table 14

Descriptive Statistics Measurement Courses Degree (N=378)

<table>
<thead>
<tr>
<th>Measurement courses</th>
<th>Bachelors</th>
<th>Masters</th>
<th>Specialist</th>
<th>Doctorate</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(25) 17.12%</td>
<td>(24) 11.21%</td>
<td>(1) 7.69%</td>
<td>(0) 0.00%</td>
<td>50</td>
</tr>
<tr>
<td>1</td>
<td>(54) 36.98%</td>
<td>(67) 31.30%</td>
<td>(2) 15.38%</td>
<td>(0) 0.00%</td>
<td>123</td>
</tr>
<tr>
<td>2-3</td>
<td>(59) 40.41%</td>
<td>(87) 40.65%</td>
<td>(6) 46.15%</td>
<td>(2) 40.00%</td>
<td>154</td>
</tr>
<tr>
<td>4-5</td>
<td>(3) 2.05%</td>
<td>(32) 14.95%</td>
<td>(1) 7.69%</td>
<td>(1) 20.00%</td>
<td>37</td>
</tr>
<tr>
<td>6+</td>
<td>(5) 3.42%</td>
<td>(4) 1.86%</td>
<td>(3) 23.07%</td>
<td>(2) 40.00%</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>(146) 100%</td>
<td>(214) 100%</td>
<td>(13) 100%</td>
<td>(5) 100.00%</td>
<td></td>
</tr>
</tbody>
</table>

Educators with a bachelor’s degree reported the lowest number of Educational Statistics coursework taken: 38.77%=0 classes, 32.65%=1 class, 26.53%=2-3 classes, 1.36%=4-5 classes and 0.68%=6 or more classes.

Educators with a specialist degree reported the highest number of Educational Statistics coursework taken: 7.14%=0 classes, 21.42%=1 class, 42.8%=2-3 classes and 14.28%=4-5 classes and 14.28%=6 or more classes. Frequencies and percentages are reported in Table 15.
Table 15

Descriptive Educational Statistics Courses Degree (N=379)

<table>
<thead>
<tr>
<th>Educational Statistics courses</th>
<th>Degree</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bachelors</td>
<td>Masters</td>
</tr>
<tr>
<td>0</td>
<td>(57)</td>
<td>(51)</td>
</tr>
<tr>
<td>38.77%</td>
<td>23.94%</td>
<td>7.14%</td>
</tr>
<tr>
<td>1</td>
<td>(48)</td>
<td>(90)</td>
</tr>
<tr>
<td>32.65%</td>
<td>42.25%</td>
<td>21.42%</td>
</tr>
<tr>
<td>2-3</td>
<td>(39)</td>
<td>(69)</td>
</tr>
<tr>
<td>26.53%</td>
<td>32.39%</td>
<td>42.85%</td>
</tr>
<tr>
<td>4-5</td>
<td>(2)</td>
<td>(1)</td>
</tr>
<tr>
<td>1.36%</td>
<td>0.47%</td>
<td>14.28%</td>
</tr>
<tr>
<td>6+</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>0.68%</td>
<td>0.94%</td>
<td>14.28%</td>
</tr>
<tr>
<td>(147)</td>
<td>(213)</td>
<td>(14)</td>
</tr>
<tr>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Educators with a bachelor's degree reported the lowest number of Math Statistics coursework taken: 42.85%=0 classes, 31.29%=1 class, 19.72%=2-3 classes, 4.08%=4-5 classes and 2.04%=6 or more classes. Educators with a specialist degree reported the highest number of Math Statistics coursework taken: 7.14%=0 classes, 35.71%=1 class, 42.85%=2-3 classes, 7.14%=4-5
classes and 7.14%=6 or more classes. Frequencies and percentages are reported in Table 16.

Table 16

**Descriptive Math Statistics Courses Degree (N=380)**

<table>
<thead>
<tr>
<th>Math Statistics courses</th>
<th>Degree</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bachelors</td>
<td>Masters</td>
<td>Specialist</td>
<td>Doctorate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>(63)</td>
<td>(72)</td>
<td>(1)</td>
<td>(2)</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td></td>
<td>42.85%</td>
<td>33.64%</td>
<td>7.14%</td>
<td>40.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(46)</td>
<td>(71)</td>
<td>(5)</td>
<td>(0)</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td></td>
<td>31.29%</td>
<td>33.17%</td>
<td>35.71%</td>
<td>0.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-3</td>
<td>(29)</td>
<td>(61)</td>
<td>(6)</td>
<td>(2)</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19.72%</td>
<td>28.50%</td>
<td>42.85%</td>
<td>40.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-5</td>
<td>(6)</td>
<td>(7)</td>
<td>(1)</td>
<td>(0)</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.08%</td>
<td>3.27%</td>
<td>7.14%</td>
<td>0.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6+</td>
<td>(3)</td>
<td>(3)</td>
<td>(1)</td>
<td>(1)</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.04%</td>
<td>1.40%</td>
<td>7.14%</td>
<td>20.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(147)</td>
<td>(214)</td>
<td>(14)</td>
<td>(5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Educators with a bachelor’s degree reported the lowest number of Research Design coursework taken: 66.66%=0 classes, 19.04%=1 class, 12.92%=2-3 classes, 0.68%=4-5 classes and 0.68%=6 or more classes.

Educators with a specialist degree reported the highest number of Research
Design coursework taken: 14.28%=0 classes, 14.28%=1 class, 50%=2-3 classes, 14.28%=4-5 classes and 7.14%=6 or more classes. Frequencies and percentages are reported in Table 17.

Table 17

*Descriptive Statistics Research Design Courses Degree (N=379)*

<table>
<thead>
<tr>
<th>Research Design courses</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bachelors</td>
</tr>
<tr>
<td>0</td>
<td>(98) 66.66%</td>
</tr>
<tr>
<td>1</td>
<td>(28) 19.04%</td>
</tr>
<tr>
<td>2-3</td>
<td>(19) 12.92%</td>
</tr>
<tr>
<td>4-5</td>
<td>(1) 0.68%</td>
</tr>
<tr>
<td>6+</td>
<td>(1) 0.68%</td>
</tr>
<tr>
<td></td>
<td>(147) 100%</td>
</tr>
</tbody>
</table>

Hypotheses Testing Results

*Hypotheses*

1. There will be no significant difference in assessment literacy on norm-referenced/criterion referenced tests, validity/reliability, and potential misuses of school personnel based on degree.
2. There will be no significant difference in assessment literacy on norm-referenced/criterion referenced tests, validity/reliability, and potential misuses based on certification.

3. There will be no significant difference in assessment literacy on norm-referenced/criterion referenced tests, validity/reliability, and potential misuses of school personnel based on years experience.

The reliability of the three separate subscales was too low to use for hypothesis testing. The theoretical differences in norm/criterion referenced subscales alpha coefficient was .533. The reliability and validity subscales alpha coefficient was .567. The potential misuses alpha coefficient subscale was .706. Two of the subscales were below the .70 that is acceptable in educational research. Since the reliability was low in two of the subscales, the researcher decided to use a composite score on the entire instrument for the dependent variable.

To test the hypotheses, the dependent variable was the overall knowledge score on the Criterion-Referenced Questionnaire. In addressing Research Hypothesis 1, 2, and 3 three one way Analysis of Variance (ANOVAs) were used to determine if there was a significant difference on performance on the Criterion-Referenced Assessment Questionnaire of school personnel based on years of experience, method of certification, and highest degree held. A significant ANOVA was followed up by a post hoc test using a Bonferroni Correction to control for Type 1 error p<.016.
The assumptions for ANOVA were examined as part of the analysis. Levene’s Test of Equality of Error was not significant $p = .236$ for years experience. However, sample sizes were unequal for years experience (0-5=90, 6-10=88, 11-15=71, 16-20=49 and 21+=82). The ANOVA for years experience was not significant, $F(4,375) = 2.41$, $p = .049$. Eta Square was .025. However, the Bonferroni Correction to control for Type 1 error criterion was .016 thus there was no significant difference between the groups. The means of the groups are reported in Table 18.

Levene’s Test of Equality of Error was not significant $p = .619$ for traditional or alternatively certified educators. Cell size was unequal in regards to type of certification (traditional=292 and alternative=88). The ANOVA for certification was not significant, $F(1,378) = 3.649$, $p = .057$. Eta Square was .009.

Levene’s Test of Equality of Error was not significant $p = .238$ for highest degree held. Cell size was unequal (bachelors=147, masters=214, and advanced degree composed of specialist and doctorate=19). The ANOVA for highest degree held was significant, $F(2,377) = 11.275$, $p < .001$. Eta square was .056. The Bonferroni Correction to control for Type 1 error revealed a significant difference in advanced degree when compared to a bachelor’s or masters both were $p < .001$.

Descriptive statistics revealed a difference between degrees as well: bachelors $M = 55.58$, $SD = 16.91$, masters $M = 56.95$, $SD = 19.62$, specialist $M = 73.78$, $SD = 16.53$ and doctorate $M = 85.20$, $SD = 9.73$. The mean of the
specialists and doctorate degrees were higher than masters and bachelor’s degrees (see Table 18).

Table 18

**Descriptive Statistics Degree/Years Experience (N=380)**

<table>
<thead>
<tr>
<th>Degree</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bachelors</td>
<td>55.58</td>
<td>16.91</td>
<td>147</td>
</tr>
<tr>
<td>Masters</td>
<td>56.95</td>
<td>19.62</td>
<td>214</td>
</tr>
<tr>
<td>Specialist</td>
<td>73.78</td>
<td>16.73</td>
<td>14</td>
</tr>
<tr>
<td>Doctorate</td>
<td>85.20</td>
<td>9.73</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>57.41</td>
<td>18.93</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5</td>
<td>57.17</td>
<td>16.85</td>
<td>90</td>
</tr>
<tr>
<td>6-10</td>
<td>58.12</td>
<td>17.70</td>
<td>88</td>
</tr>
<tr>
<td>11-15</td>
<td>60.38</td>
<td>20.28</td>
<td>71</td>
</tr>
<tr>
<td>16-20</td>
<td>60.85</td>
<td>17.47</td>
<td>49</td>
</tr>
<tr>
<td>21+</td>
<td>52.29</td>
<td>21.21</td>
<td>82</td>
</tr>
<tr>
<td>Total</td>
<td>57.41</td>
<td>18.93</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Certification</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>56.40</td>
<td>18.63</td>
<td>292</td>
</tr>
<tr>
<td>Alternative</td>
<td>60.78</td>
<td>19.64</td>
<td>88</td>
</tr>
<tr>
<td>Total</td>
<td>57.41</td>
<td>18.93</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER V

SUMMARY

Conclusions

With the increased importance of assessment since the passage of No Child Left Behind, it has become very important that educators are assessment literate. A review of related literature has shown that teachers and administrators spend one-third to one-half of their time on assessment activities (Plake, 1993; Stiggins, 1988, Stiggins, 1991). The literature also revealed that teachers and administrators may lack the training to effectively perform their role in the assessment process required by NCLB (Hollenbeck, Tindal, & Almond, 1998; O'Sullivan & Chalnick, 1991; Plake, 1993; Popham, 2006c; Stiggins, 1995). What type of training have educators received in order to help prepare them to understand the assessment process? This study collected data in order to help determine what type of classes educators have taken in measurement, assessment, and statistics.

The data collected on the number of measurement courses taken by teachers, counselors, and administrators demonstrated an increase in classes taken. Apparently, universities are requiring more measurement classes for teachers, counselors, and administrators. Eighty-six percent of teachers reported taking one or more classes in measurement. Seventy percent of counselors reported taking at least one or more classes in measurement. Ninety-four percent of administrators reported taking at least one or more measurement classes. The review of literature documented that only 20% of the
state required coursework in assessment and only 21% of teachers and 16% of administrator’s degree programs required a course in assessment (Stiggins, 1991; Wolmut, 1988). All three groups reported taking more coursework in measurement than previously reported. An even more interesting trend was revealed in the data; 54% of the participants reported taking two or more classes. An earlier study reported that only 1.4% of universities required two or more classes in order to graduate (Roder, 1972). Educators and universities are recognizing the increased importance of measurement and are offering more coursework in this area over what had been previously reported in literature.

The data collected on the number of educational statistics courses taken revealed a similar trend. Seventy-one percent of the participants reported taking at least one class in educational statistics. Thirty-three percent of the participants reported taking two or more classes in educational statistics. All groups reported taking more coursework than had been previously reported. The data collected on math statistics coursework also demonstrated an increase in classes taken by the participants. Sixty-four percent of the participants reported taking at least one class in math statistics. Thirty-one percent of the participants reported taking two or more classes in math statistics. The participants again reported a higher number of classes taken than reported previously.

The data collected on research design classes also showed an increase in classes taken over that previously reported. Fifty percent of the participants reported at least one class in research design. Twenty-three percent of the participants reported taking two or more classes in research design. Research
design classes had a smaller increase in the number of classes taken when compared to other areas in the study. Popham (2007) emphasized the importance of understanding research design and the successful implementation of the school improvement process. These classes are most commonly offered, at the specialist and doctorate level. Since educators with a specialist or doctorate were a small percentage of the population of the study, it would make intuitive sense that there would be a smaller increase because a smaller number of participants would have coursework in these areas.

Coursework in these four areas were also analyzed based on type of certification; traditional or alternative. Since the participants were the same, similar trends showed up in the data. All the courses, educational measurement, educational statistics, math statistics, and research design, showed an increase in coursework taken over previously reported data. However, there were several trends that were noted in the data. Traditionally certified teachers reported taking more classes in educational measurement and educational statistics. Alternatively certified educators reported taking more math statistics and research design classes. Educational measurement and educational statistics classes are offered within the college of education. Alternatively certified teachers graduated with degrees outside the college of education, therefore, it would make sense that they would not have taken as many educational measurement and statistics classes. It also would make sense that alternatively certified teachers would have taken more math classes that would have been offered outside the college of education.
Data were also collected on the number of courses taken by educators in educational measurement, educational statistics, math statistics, and research design taken based on highest degree completed. Eighty-three percent of educators with a bachelor’s degree reported taking one or more educational measurement classes. Eighty-nine percent of educators with a master’s degree reported taking one or more educational measurement class. Ninety-two percent of educators with a specialist degree reported taking one or more educational measurement class. Educators with a doctorate degree reported the highest number of educational measurement classes taken. One hundred percent of educators with a doctorate degree reported taking two or more classes. The number of classes reported by the participants was well above that reported in earlier studies. Schafer and Lissetz (1987) reported the percentage of degrees that required at least one measurement class to be, “57% for bachelor’s programs, 69% for master’s and 70% for doctoral programs” (p. 62). Universities may be requiring more coursework in educational measurement. However, based on the low performance on the Criterion-Referenced Questionnaire by participants, this increased coursework may not be effectively meeting the needs of educators.

This trend continued when the data were analyzed for educational statistics classes. Sixty-one percent of educators reported taking one or more educational statistics classes. Seventy-six percent of educators with master’s degrees reported taking one or more educational statistics courses. Ninety-two percent of educators with a specialist degree reported taking one or more
classes. Eighty percent of educators with a doctorate reported taking two or more classes in educational statistics. All groups reported a higher number of classes taken than had been reported in previous studies.

Data were also analyzed on the number of research design classes taken by the participants. Thirty-three percent of educators with a bachelor's degree reported taking one or more classes in research design. Fifty-eight percent of educators with a master's degree reported taking one or more classes in research design. Eighty-five percent of educators with a specialist degree reported taking one or more classes in research design. Eighty percent of educators with a doctorate degree reported taking one or more research design classes. According to the survey, research design courses are mostly limited to those with advanced degrees. These seem to make intuitive sense because research design classes are most often taken during an advanced degree program such as a specialist and doctorate degree. Schafer and Lissitz (1987) noted that advanced degrees required more coursework in measurement and assessment. Mayo (1967) in an earlier study reported more measurement, assessment, and statistics were required at the graduate level in education degree programs. Specialist and doctoral students take more measurement, assessment, and statistic coursework in order to help prepare them to complete the dissertation process in advanced degree programs. However, this same coursework may not meet the needs of practicing educators interpreting data in order to complete the school improvement process. Specialized classes may
need to be developed in order to help educators interpret data in order to complete their school improvement plan as they work toward AYP.

Popham (2007) expressed concern that educators need more knowledge in the area of research design. Popham noted, “Before placing confidence in empirical investigations, especially in studies that may influence the way we educate children, we need to be certain the researchers adhered to the fundamental canons of research design” (p. 88). Thus, it is important that educators understand research design so they help ensure that data has been gathered and analyzed correctly. Major changes in educational policy should only be based on studies that adhere to basic rules of research design. The participants of the study reported taking more coursework in all areas of assessment, measurement, and statistics than had been previously reported. Educators with a specialist and doctorate degree reported taking more classes than the other groups. The data seemed to support that there had been an increase in the number of assessment, measurement, and statistics classes taken by educators in coastal Alabama and Mississippi. It appears that universities in coastal Alabama and Mississippi are requiring more coursework in assessment, measurement, and statistics in order to complete a degree in the field of education since NCLB.

Years of service was assessed to determine if experience had a significant impact on a participant’s performance of the Criterion-Referenced Questionnaire. The ANOVA was non-significant for years of experience. This result is in contrast to previous studies. Earlier research reported most of what
teachers know about measurement is learned through trial and error experience in the classroom (Wise et al., 1991). Similarly, Stiggins (1988) reported that teachers cite their experience in the classroom as their most influential source of measurement information. A study in 2005 reported that inservice teachers were significantly more assessment literate than preservice teachers (Mertler).

Mertler’s research instrument, Classroom Assessment Literacy (CALI), focused on scoring, administering, interpreting, and using test results. The Criterion-Reference Questionnaire knowledge items used in this study focused on theoretical understanding of test, reliability, validity, and potential misuses of tests. It is important to remember that performance on the Criterion-Referenced Questionnaire was judged by an overall composite score on the questionnaire and not broken down into subscales. However, educators, regardless of years experience, scored very similarly on the Criterion-Referenced Questionnaire. The mean, based on years experience, only differed by eight points: 60.85% and 52.29%. Educators with over 21 years experience scored lower than all the other groups. The findings of this study stand in stark contrast to earlier studies. Perhaps this is the result of the school improvement process in which educators are being asked to analyze data at the school district and individual school level. In previous years, teachers were primarily concerned with interpreting and applying their own teacher-made tests in their classrooms. The school improvement process requires a more global understanding of assessment and its uses. Teachers only have classroom experience to apply to these new situations. Inservice training may be needed in order to help educators make the
successful transition from classroom assessment to building and district level. Further research would be needed to determine which areas of assessment, measurement, and statistics are influenced by experience. Mertler (2005) made a similar recommendation that further research was needed to determine which assessment skills are best learned on the job.

This study also analyzed performance on the Criterion-Referenced Questionnaire based on the job of the participants. The mean of teachers and administrators only differ by three percentage points: 57.022% and 60.16%. The lowest mean reported by both groups was likely the result of insufficient training in assessment, measurement, and statistics. Previous research has reported that only about half of teacher education programs require a course in assessment or measurement (Schafer, 1991). Schafer and Lissitz (1987) found that only 15% of administrator training required a class in assessment or measurement. Administrators do not receive significantly more training in assessment, measurement, and statistics. Thus, it would make sense that teachers and administrators would perform similarly on the Criterion-Referenced Questionnaire. However, counselors had a higher mean than teachers and administrators at 66.30%. The higher performance of counselors was most likely the result of the job experience. The fact that on the job experience could have a positive impact on a person’s knowledge of assessment and measurement was well documented by the literature (Mertler, 2005; Stiggins, 1988; Wise et al, 1991). Based on the researcher’s experience, counselors handle more standardized test data than any other employee in a school building thus leading
to a better understanding of assessment and measurement and this assumption is supported by the findings of this study.

Educators’ knowledge of measurement, assessment, and statistics were examined using twenty-four multiple choice questions. As reported previously, the participants scored relatively low on the Criterion-Referenced Questionnaire. The participants answered only 47.36% of the questions correctly that were based on the theoretical differences of norm and criterion-referenced tests. Only 53.8% of the participants could correctly identify the theoretical base for norm-referenced test. Even fewer participants (29.9%) could identify that variability around a cut score is the basis for criterion-referenced testing. Most participants did not understand random error in relation to classical test theories. The findings of this study seemed to support earlier research that shows teachers and administrators seem to have substantial gaps in their knowledge of statistics and assessment (Gullickson, 1986; Popham, 2006b; Popham, 2006c; Stiggins, 1998; Zwick et al., 2008). Zwick's study specifically mentions that teachers and administrators do not understand measurement error.

Educators scored low on the questions regarding validity and reliability. Only 43.3% of the participants could correctly identify the main types of validity. The participants had even more trouble identifying the types of reliabilities; only 26.2% could identify the main types of reliability. The participants’ low performance on these questions seems to support previous research. Plake (1993) noted that only 13% of teachers could correctly answer her questions on reliability. A more recent study noted that educators did not understand the
concept of reliability (Zwick et al., 2008). Several other studies have documented that educators have a poor understanding of validity (Hollenbeck, Tindal, & Almond, 1998; Plake, 1993; Popham, 2006b; Zwick et al., 2008). Educators’ lack of knowledge of validity and reliability needs to be addressed. Research has shown that reliability and validity of test scores are influenced by their interpretation (Hollenbeck et al.; Popham). Thus, educator knowledge of validity and reliability influences all the test data gathered during the school improvement process.

The participants scored a little higher on questions regarding potential misuses of assessments. Most participants could correctly answer that test validity and reliability are reduced by inappropriate test administration and preparation practices. However, most of the participants could not identify that systematic error is increased by inappropriate test preparation and administration practices. Thus, the majority of educators did not understand why inappropriate test administration and preparation create a systematic error in an educational measurement. Teacher and administrator failure to understand systematic error is documented in the literature. Researchers have noted that teachers did not understand that inappropriate test accommodation and modification create a systematic error in educational test data (Hollenbeck et al., 1998). Plake (1993) noted that teachers had trouble answering questions on unethical or misuses of assessment. The findings of this study seem to support earlier research that teachers have a poor understanding of the ways assessment are misused.
The complete analysis of the knowledge items on the Criterion-Referenced Questionnaire seems to document that educators may not be literate in measurement, assessment, and statistics. The participants scored low on the Criterion-Referenced Questionnaire. In addition, the participants scored low on several individual questions based on theoretical differences in norm and criterion-referenced tests, reliability and validity, and potential misuses and further supports that educators may not be assessment literate. Although more training in measurement, assessment, and statistics was reported by the participants, this did not necessarily increase their assessment literacy. This supports previous research that educators may lack the training to be assessment literate (Hollenbeck et al., 1998; O’Sullivan & Chalnick, 1991; Plake, 1993; Popham, 2006b; Stiggins, 1995; Zwick et al., 2008).

An examination of the Likert items revealed some interesting trends. More participants had a favorable opinion of educational statistics than had a negative opinion. This is in stark contrast to earlier research that found teachers had a strong bias and dislike for educational statistics (Gullickson, 1986; Mayo, 1967). Mayo reported that educators loathe statistical concepts. Perhaps the increased favorability of educational statistics is the result of their increased use due to the passage of NCLB.

A majority of the participants reported that they agreed or strongly agreed that the statistical analysis of test data was useful to them as a teacher or administrator. This is in contrast to earlier studies that showed teachers and administrators thought statistical analysis of test data was more work than the
analysis was worth (Gullickson, 1986; Stiggins, 1998; Lai & Waltman, 2008). Only 27% of the participants agreed or strongly agreed with the statement that the school improvement process could be effectively implemented without a detailed statistical analysis of test data. Thus, a majority of educators in this study believed the statistical analysis of data was an important part of the school improvement process. This could be the result of studies that have begun to show that improving the assessment literacy of teachers leads to an increase in student achievement because educators are familiar with research that ties improved assessment literacy to student achievement (Black & William, 1998; Meisels, Atkins, Xue & Bickel, 2003; Rodriguez, 2004).

The Likert items revealed that many educators did not believe that they had received adequate training in assessment, measurement, and statistics. Forty-one percent of the participants disagreed or strongly disagreed with the statement, “I have received adequate training in assessment, measurement, and statistics,” to carry out their job. In addition, 43.5% disagreed or strongly disagreed with the statement that their school districts provided inservice training in assessment, measurement, and statistics. Forty-six percent reported they had not been informed that there were Standards for Educational and Psychological Testing. The data seemed to show that many educators were dissatisfied with their training in assessment, measurement, and statistics. Earlier research documented that educators felt unprepared to carry out their job in the areas of assessment, measurement, and statistics (Mertler, 2005; Plake, 1993; Popham, 2006c). Researchers even reported a similar percentage of teachers believed
themselves unprepared at 47% (Wise et al., 1991). The data collected by this study seems to support the previous studies that found many teachers feel inadequately prepared in the area of assessment, measurement, and statistics.

Limitations

The study had several limitations. Time was one of the major limitations with the study. The research project began with IRB approval on January 31, 2010 and continued until April 14, 2010, so schools were only surveyed over a two and a half month period of time. The study would have been more thorough if data could have been gathered over an entire school year.

An additional limitation with the study was the resistance of the participants. The Criterion-Referenced Questionnaire contained twenty-four knowledge based questions. Many of the participants verbally expressed resentfulness in being asked to answer anything that assessed their personal knowledge. This hostility, in many cases, was openly expressed at faculty meetings while the instrument was being administered. This behavior was observed numerous times by the researcher as well as some principals who agreed to distribute the questionnaire and return it to the researcher. Educator’s defensiveness to the questionnaire was probably a result in their lack of training in measurement, assessment, and statistics. Many of the schools that agreed to take part in the study asked the researcher to distribute the questionnaire, in part, because of their faculty being unreceptive. Principals, being unwilling to distribute and collect the questionnaire, proved to be another limitation in the study. Therefore, this limited the number of schools that could be surveyed.
Another limitation was the number of school districts that agreed to take part in the study. There were eleven school districts in the coastal counties of Mississippi. Only two school districts agreed to take part in the study. Twelve schools in Mississippi were surveyed. In Alabama, three school districts were within the coastal counties. Two of the three school districts in Alabama agreed to take part in the study. A total of nine schools were surveyed in Alabama. Thus, the number of school districts and schools surveyed was a limitation of the study.

The participants in this study were practicing educators that included teachers, principals, assistant principals, and counselors within the coastal areas of the states of Alabama and Mississippi. Including all districts in Alabama and Mississippi would have made the findings of this study more generalizable to a larger population. Any inferences in the findings of this study to a larger population should consider these limitations.

Recommendations for Policy

The research shows that teachers and administrators do not receive enough quality training in assessment, measurement, and statistics to be assessment literate. The data gathered during this research project seemed to support previous studies that demonstrated that teachers and administrators seem to have gaps in their knowledge of measurement, assessment, and statistical reasoning. Educators' experience did not increase their performance on the instrument. However, educators that had advanced degrees performed significantly better on the Criterion-Referenced Questionnaire. It is important to remember that advanced degree in this study included educators with a
specialist or doctoral degree. Educators that had a bachelors or masters degree had similar performance on the instrument.

Doctoral and specialist degree programs seemed to better prepare educators to understand measurement, assessment, and statistics. This finding would seem to reflect the fact that advanced degrees require more coursework in measurement, assessment, and statistics. This coursework is needed to prepare a student to gather and analyze data in order to complete the dissertation process. However, this revealed a problem that needs to be addressed. Based on the researcher’s experience, most educators do not hold an advanced degree (specialist and doctoral). Therefore, if educators are to be more assessment literate, more coursework in measurement, assessment, and statistics should be required at the bachelors and masters level. This could be accomplished in several ways. First, the courses from the specialist to doctoral level could be moved down and offered in master level programs. This suggestion would seem not to be the most practical since measurement, assessment, and statistical courses offered at the doctoral level have prerequisites. These prerequisite courses help to ensure the student has the basic knowledge in order to understand advanced concepts in measurement, assessment, and statistics. Simply moving these courses to a master’s level program without students having prerequisite courses might cause more confusion than any benefit they could provide. This researcher would recommend that more coursework be required at the bachelor’s and master’s levels.
Second, new courses could be developed to address the need of increasing the assessment literacy of educators with a bachelors or masters degree. These new courses should focus on training educators how to gather and analyze educational data at the school and district level. The coursework should be designed to help educators effectively implement data analysis in order to complete the school improvement process required by NCLB. A third way to increase measurement, assessment, and statistical reasoning of educators would be to provide inservice training in this area. An excellent example of this type of inservice would be the “STARS” program that is currently being implemented in Nebraska with the help of Barbara Plake and Richard Stiggins. Any type of inservice training provided should help teachers understand concepts, such as: reliability and validity, theoretical foundation of testing, and the potential harmful effects when testing is misused. Additional inservice training should be provided to help educators interpret, analyze, and apply their knowledge of measurement, assessment, and statistics to the school improvement process. This would seem to be the most practical method in order to increase assessment literacy among educators. However, in this method, there is also a problem. How many educators within a school district are qualified to lead an inservice training in measurement, assessment, and statistics? A reasonable way to address this issue would be for school districts and local universities to form a partnership. Measurement, assessment, and statistics experts at the universities could help develop and lead inservice training for a school district. Another practical extension of this would be for
measurement experts at the university to train a cohort of educators that could lead the inservice training within the school districts.

This lack of training raises a simple question: How do we insure that educators receive adequate training in measurement? Researchers recommended that teacher and administrator training programs increase the number of courses in measurement and statistics in order to complete the program (Wise & Lukin, 1993). In addition, colleges and universities could require competency testing in the area of measurement, assessment, and statistics as part of their comprehensive exit exam for degree completion. Higher education departments could tailor these exit exams to meet the needs of their individual institutions and states. Another recommendation is that state certification agencies could simply rewrite licensing requirements for teachers and administrators requiring more coursework in measurement and statistics (Wise & Lukin). Changing the certification requirements would help to ensure that educators are more assessment literate in the future.

Recommendations for Future Research

The results of this study revealed that the type of degree held by the individual significantly affected his/her performance on the Criterion-Referenced Questionnaire. Future research should be conducted to help determine the types of classes that would most likely improve the assessment literacy of educators. Other researchers have made similar recommendations. Carter (1984) reported, “The scope and sequence in typical preservice measurement courses should be reexamined in light of teachers in security about the strength of their background
in testing” (p. 60). Mertler (2005) noted preservice training in assessment would be improved if the skills and knowledge needed by teachers could be identified.

Results of this study seemed to demonstrate that educators did not feel adequately prepared in assessment, measurement, and statistics. Further research is recommended in the area of inservice training of assessment, measurement and statistics. Research should be conducted to determine the types of inservice training offered to educators. The content of these inservice training programs should be analyzed as well.
APPENDIX A

INSTRUMENT

Criterion-Referenced Questionnaire

Demographics: Please Circle One:
1. Certification: Traditional Alternative
2. Highest Degree Earned: Bachelors Masters Specialist Doctorate
3. Gender: Male Female
4. Grade Level: K - 5th 6th - 8th 9th - 12th
5. Job Title: teacher counselor administrator
6. Years of Service: 0 - 5 6 - 10 11 - 15 16 - 20 21+
7. Number of Courses Taken in Educational Measurement and Assessment:
   None 1 2 – 3 4 - 5 6+
8. Number of Courses Taken in Educational Statistics:
   None 1 2 – 3 4 – 5 6+
9. Number of Courses Taken in Math Statistics:
   None 1 2 – 3 4 – 5 6+
10. Number of Courses Taken in Research Design:
    None 1 2 – 3 4 – 5 6+

Directions: Please Circle One:
1. Variability around the mean is the theoretical basis for which type test?
   A. Criterion-referenced
   B. Norm-referenced
   C. Ability (Aptitude)
   D. Performance
2. Students’ performance on ACT and SAT are used by universities as predictors of student success in college. The predictive ability of these tests is an example of what type of validity?
   A. Content
   B. Construct
   C. Criterion
   D. Aptitude
3. Raw scores and cut-scores are commonly reported for which type of test?
   A. Criterion-referenced
4. Variability around a cut-score is the main theoretical basis for which type of test?
   A. Criterion-referenced
   B. Norm-referenced
   C. Aptitude
   D. Performance

5. Measures of central tendency such as mean, median, mode, and standard deviation would be commonly reported for which type of test?
   A. Criterion-referenced
   B. Norm-referenced
   C. Aptitude
   D. Performance

6. What type of test is best suited in determining an individual performance in relation to a larger group?
   A. Criterion-reference
   B. Norm-referenced
   C. Aptitude
   D. Performance

7. What type of test is commonly used to determine mastery learning of objectives within an educational domain?
   A. Criterion-referenced
   B. Norm-referenced
   C. Aptitude
   D. ASVAB

8. Identify the main types of validity?
   A. Test-retest, equivalent form, split-half, inter-rater
   B. Criterion, content, construct, face
   C. Test-retest, equivalent form, criterion, face
   D. Inter-rater, split-half, face, concurrent

9. Identify the main types of reliabilities:
   A. Test-retest, equivalent form, split-half, inter-rater
   B. Criterion, content, construct, face
   C. Test-retest, equivalent form, criterion, face
   D. Inter-rater, split-half, face, concurrent

10. The bell curve is most closely linked with what type of test?
    A. Criterion-referenced
    B. Norm-referenced
11. Reliability is an important measure of a test or set of test scores. Which of the following represents an acceptable measure of reliability:
   A. .20
   B. .80
   C. .60
   D. .50

12. In order for any test to have meaningful scores, the test must be:
   A. Short and easy to grade
   B. Long and easy to grade
   C. Short and difficult
   D. Valid and reliable

13. Inappropriate test administration and preparation practices cause problems in standardized testing because:
   A. Reduces confusion of students
   B. Reduces confusion of teachers
   C. Reduces error
   D. Reduces validity and reliability

14. Most educational experts would set mastery of educational objectives within a domain at:
   A. 40%
   B. 55%
   C. 60%
   D. 78%

15. The validity and reliability of tests may be jeopardized by which of the following:
   A. Poor test administration
   B. Inappropriate test preparation
   C. Systematic error
   D. All of the above

16. Inappropriate test preparation and administration practices can increase what type of error?
   A. Systematic error
   B. Random error
   C. Normal error
   D. Criterion error

17. NCLB requires that school/school districts use only test scores that are __________ in the school improvement process.
A. Simple and easy to read
B. Reliable and valid
C. Percentage and numerical
D. Difficult and time consuming

18. Increasing the systematic error in any measurement:
   A. Increases the accuracy
   B. Reduces the error
   C. Increases the reliability and validity
   D. Decreases the reliability and validity

19. In general, classical test theories are designed to address what type of error?
   A. General error
   B. Random error
   C. Systematic error
   D. Index error

20. When a test is used in a manner that harms students and teachers or for a purpose that the test was not developed, that is called:
   A. Test expansion
   B. Test misuse
   C. Test abuse
   D. Data gathering

21. It is inappropriate to use a student test performance to evaluate a teacher because of which of the following:
   A. Student population varies between classrooms
   B. The social economic status varies between communities
   C. The experience of the teacher varies between schools
   D. Validity and reliability of tests has not been established for this purpose

22. It is inappropriate to use student test performance to evaluate a school because:
   A. Student population varies between classrooms
   B. The social economic status varies between communities
   C. The experience of the teacher varies between schools
   D. Validity and reliability of tests has not been established for this purpose

23. If teachers do not read test directions to students completely during standardized tests, this may result in:
   A. Decreasing test taking time
   B. Reducing validity and reliability of the test
   C. Increasing test scores
D. Increasing test anxiety

24. If teachers do not follow all test accommodations and modification for special education students, this may result in:
   A. Decreasing test taking time
   B. Reducing validity and reliability of the test
   C. Increasing test scores
   D. Increasing test anxiety

Please circle the response that best reflects your opinion using the following scale. The following terms have been used: Strongly Disagree (SD=1); Disagree (D=2); Not Sure (NS=3); Agree (A=4); Strongly Agree (SA=5).

25. I have a favorable opinion of educational statistics.  
SD D NS A SA
1 2 3 4 5

26. I think statistical analysis of test data is useful to me as a teacher or administrator.  
SD D NS A SA
1 2 3 4 5

27. I have received adequate training in measurement, assessment and statistics to effectively analyze test data as part of my job.  
SD D NS A SA
1 2 3 4 5

28. I believe the school improvement process can be effectively implemented without the detailed statistical analysis of test data.  
SD D NS A SA
1 2 3 4 5

29. The school improvement process should be based on test data gathered by assessments given by the school district.  
SD D NS A SA
1 2 3 4 5

30. I believe the tests my school/school district uses are reliable and valid.  
SD D NS A SA
1 2 3 4 5

31. My school/school district clearly
state to teachers and administrators how the validity and reliability was established for tests it administers to students.  

32. My school/school district offers inservice training on educational statistics, measurement, and assessment for teachers and administrators.  

33. The use of student test scores to evaluate teachers’ performance is a valid use of these tests.  

34. There are national standards for educational measurement that are stated in the Standards for Educational and Psychological Testing (1999). Standards have been discussed and explained to me as part of training as a teacher or administrator.
APPENDIX B

IRB APPROVAL

THE UNIVERSITY OF SOUTHERN MISSISSIPPI

Institutional Review Board

118 College Drive #5147
Hattiesburg, MS 39406-0001
Tel: 601.266.6820
Fax: 601.266.5509
www.usm.edu/irb

HUMAN SUBJECTS PROTECTION REVIEW COMMITTEE
NOTICE OF COMMITTEE ACTION

The project has been reviewed by The University of Southern Mississippi Human Subjects Protection Review Committee in accordance with Federal Drug Administration regulations (21 CFR 26, 111), Department of Health and Human Services (45 CFR Part 46), and university guidelines to ensure adherence to the following criteria:

- The risks to subjects are minimized.
- The risks to subjects are reasonable in relation to the anticipated benefits.
- The selection of subjects is equitable.
- Informed consent is adequate and appropriately documented.
- Where appropriate, the research plan makes adequate provisions for monitoring the data collected to ensure the safety of the subjects.
- Where appropriate, there are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of all data.
- Appropriate additional safeguards have been included to protect vulnerable subjects.
- Any unanticipated, serious, or continuing problems encountered regarding risks to subjects must be reported immediately, but not later than 10 days following the event. This should be reported to the IRB Office via the “Adverse Effect Report Form”.
- If approved, the maximum period of approval is limited to twelve months. Projects that exceed this period must submit an application for renewal or continuation.

PROTOCOL NUMBER: 10012704
PROJECT TITLE: Criterion-Referenced Assessment Literacy of Educators
PROPOSED PROJECT DATES: 01/30/10 to 01/30/11
PROJECT TYPE: Dissertation or Thesis
PRINCIPAL INVESTIGATORS: James David King
COLLEGE/DIVISION: College of Education & Psychology
DEPARTMENT: Educational Leadership
FUNDING AGENCY: N/A
HSPRC COMMITTEE ACTION: Expedited Review Approval
PERIOD OF APPROVAL: 01/28/10 to 01/27/11

[Signature]
Lawrence A. Hosman, Ph.D.
HSPRC Chair
APPENDIX E
SUPERINTENDENT CONTACT LETTER

Dear Superintendent,

I would like to introduce myself. My name is James David King. As a fellow educator, I am interested in our training and knowledge of assessment, measurement, and statistical reasoning because of the increased importance of data analysis required by the “No Child Left Behind Act” and the school improvement process. I have been a practicing educator for twenty-three years in the Mobile County Public School System. I am in the process of completing my doctorate degree in educational leadership and research at the University of Southern Mississippi.

Your school district is being asked to take part in a study of the criterion assessment questionnaire. The purpose of this study is to gather data concerning the training and knowledge of educators in assessment, measurement, and statistical reasoning. After the completion of the study, I would be happy to make a presentation to your school district on the results of research. In addition, I would be willing to provide a free inservice to your district to address needs, as determined by the study.

I have attached a copy of my research questions and questionnaire for you to review. The attached questionnaire covers four main issues related to assessment and measurement as well as basic demographic variables. Completion of the questionnaire should take no more than 10-15 minutes for each teacher/administrator taking part in the study.

Data will be aggregated and a summary will be submitted as part of completing the dissertation process at the University of Southern Mississippi and may be presented in a professional venue. No individual school or school district will be identified in the summary report. Upon completion of data compilation, all questionnaires will be destroyed. Any information inadvertently obtained during the course of this study will remain completely confidential.

Participation in this project is completely voluntary. If you will allow me to survey your school district, I will need permission via e-mail at jking01@centurytel.net or a letter with your district’s letterhead. If you have any questions concerning this research, please contact James David King at 251-865-1233 or 251-508-2552. This research is being conducted under the supervision of Thelma Roberson, Ph.D. Thelma.Roberson@usm.edu.

Thank you for your consideration and help with this project.

Sincerely,
James David King, Ed.S.
13900 Tom Gaston Rd.
Mobile, AL 36695

This project has been reviewed by the Institutional Review Board, which ensures that research projects involving human subjects follow federal regulations. Any questions or concerns about rights as a research participant should be directed to the Chair of the Institutional Review Board, the University of Southern Mississippi, 118 College Drive #5147, Hattiesburg, MS 39406-0001, (601)266-6820.
Dear Principal,

I would like to introduce myself. My name is James David King. As a fellow educator, I am interested in our training and knowledge of assessment, measurement, and statistical reasoning because of the increased importance of data analysis required by the “No Child Left Behind Act” and the school improvement process. I have been a practicing educator for twenty-three years in the Mobile County Public School System. I am in the process of completing my doctorate degree in educational leadership and research at the University of Southern Mississippi.

I have been given permission by your superintendent for you to take part in a study of the criterion assessment questionnaire. The purpose of this study is to gather data concerning the training and knowledge of educators in assessment, measurement, and statistical reasoning.

Enclosed are my research questions and questionnaire. The questionnaire covers four main issues related to assessment and measurement as well as basic demographic variables. Completion of the questionnaire should take no more than 10-15 minutes for each teacher/administrator taking part in the study.

Data will be aggregated and a summary will be submitted as part of completing the dissertation process at the University of Southern Mississippi and may be presented in a professional venue. No individual school or school district will be identified in the summary report. Upon completion of data compilation, all questionnaires will be destroyed. Any information inadvertently obtained during the course of this study will remain completely confidential.

Participation in this project is completely voluntary. Please return all surveys to: James D. King 13900 Tom Gaston Rd., Mobile, AL 36695. A postage paid box has been attached for your convenience. If you have any questions concerning this research, please contact James David King at 251-865-1233 or 251-508-2552, or e-mail me at jking01@centurytel.net. This research is being conducted under the supervision of Thelma Roberson, Ph.D. Thelma.Roberson@usm.edu.

Thank you for your consideration and help with this project.

Sincerely,
James David King, Ed.S.
13900 Tom Gaston Rd.
Mobile, AL 36695

This project has been reviewed by the Institutional Review Board, which ensures that research projects involving human subjects follow federal regulations. Any questions or concerns about rights as a research participant should be directed to the Chair of the Institutional Review Board, the University of Southern Mississippi, 118 College Drive #5147, Hattiesburg, MS 39406-0001, (601)266-6820.
APPENDIX G

INFORMED CONSENT LETTER

Dear Educator,

You are being asked to take part in a pilot study of the criterion assessment questionnaire. The purpose of this study is to gather data concerning the training and knowledge of educators in assessment, measurement, and statistical reasoning. As a fellow educator, I am interested in our training and knowledge of assessment, measurement, and statistical reasoning because of the increased importance of data analysis required by the “No Child Left Behind Act” and the school improvement process.

The attached questionnaire covers four main issues related to assessment and measurement as well as basic demographic variables. Completion of the questionnaire should take no more than 10-15 minutes. Please do not put your name or any other identifying information on the questionnaire.

Data will be aggregated and a summary will be submitted as part of completing the dissertation process at the University of Southern Mississippi and may be presented in a profession venue. No individual school or school district will be identified in the summary report. Upon completion of data compilation, all questionnaires will be destroyed. Any information inadvertently obtained during the course of this study will remain completely confidential.

Participation in this project is completely voluntary. Please feel free to decline participation at any point without concern over penalty, prejudice, or any other negative consequence. If you have any questions concerning this research, please contact James David King at 251-865-1233 or 251-508-2552. This research is being conducted under the supervision of Thelma Roberson, PhD. Thelma.Roberson@usm.edu.

By completing and returning the attached questionnaire you are giving your permission for this anonymous and confidential data to be used for the purpose described above.

Thank you for your consideration and help with this project.

Sincerely,

James David King, Ed.S
This project has been reviewed by the Institutional Review Board, which ensures that research projects involving human subjects follow federal regulations. Any questions or concerns about rights as a research participant should be directed to the Chair of the Institutional Review Board, the University of Southern Mississippi, 118 College Drive #5147, Hattiesburg, Ms 39406-0001, (601)266-6820.
REFERENCES


Cronbach, L. J. (1964). Course improvement through evaluation. *Teacher College Record, 64,* 672-683.


*Educational Leadership*. 63(50), 82-83.

*Educational Leadership*. 63(60), 84-85.

*Educational Leadership*, 64(2), 90-91.


Popham, W. J. (2007). A test is a test is a test – not.  
*Educational Leadership*. 64(4), 88-89.


*Phi Delta Kappan*, 81(3), 204-208.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS.  

Roeder, H. H. (1972). Are today’s teachers prepared to use tests?  

*Chicago Sun-Times*.


