2008

A Study Of An Attempt To Improve The Reliability Of Teachers' Holistic Scores Of Elementary Writing Through In-house Profess

Lisa Farmer
University of Central Florida

A STUDY OF AN ATTEMPT TO IMPROVE THE RELIABILITY OF TEACHERS' HOLISTIC SCORES OF ELEMENTARY WRITING THROUGH IN-HOUSE PROFESSIONAL DEVELOPMENT

by

LISA EPPS FARMER

B.S. Xavier University of Louisiana, 1977
M.A. University of Central Florida, 1993

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Education in Curriculum and Instruction
in the Department of Teaching and Learning Principles
in the College of Education
at the University of Central Florida
in Orlando, Florida

Summer Term
2008

Major Professors: Michael Hynes
Susan Wegmann

ABSTRACT

This study evaluated the effectiveness of a school-based training that attempted to improve the reliability of holistic scores teachers assigned to the writings of elementary school students. Seventeen teachers at one suburban elementary school located in the Southeastern United States participated in three training sessions that allowed for scoring practice and group discussions. The trainers, or presenters, were "faculty-experts." A comparison of scores the participants assigned to students' writings before and after the training was conducted. The analyses included *t*-tests that compared the participants' mean scores to the scores assigned by raters from the state, a within-group analysis of reliability as measured by Cronbach's Alpha, and percentage agreement analyses. The results suggested that the in-house training activities promoted higher inter-rater reliability of scores assigned to students' writings by the teachers in this study.

This study also compared teachers identified as being highly confident writers with teachers who reported low levels of self-confidence related to writing. Prior to the training, the highly confident teachers' scores tended to be lower than the state scores and the scores assigned by their less confident peers. During group discussions, however, the "high-confidence" group was just as likely to change their scores to a higher level as to a lower level, and by the end of the training, both groups demonstrated more consistent score patterns.

Dedicated to the memory of Mother Dear, Lorraine Blache Epps,
who taught me to never give up and to dream my dreams.

Dedicated to my husband, Gerald Eric Farmer,
who encouraged me to pursue my personal goals.

Dedicated to my aunt, Amelia B. Barnes, who often recounted
her teaching experiences and first sparked my interest
in the teaching profession.

ACKNOWLEDGMENTS

My gratitude is expressed to my co-chairs, Dr. Michael C. Hynes and Dr. Susan J. Wegmann, and my committee members, Dr. Jeffrey Kaplan, Dr. Nance S. Wilson, and Dr. Cynthia J. Hutchinson. Thanks for sharing your expertise and for your guidance and support during this dissertation process.

Dr. Hynes, thanks for taking me under your wing from the start. Your calm encouragement helped me put my doubts and frustrations into perspective.

Dr. Wegmann, thanks for the hours you spent reading the earliest drafts of this dissertation. I valued every bit of your advice and know that without you, my journey would have been arduous indeed.

Thanks to the principal of the school in this study who allowed me to orchestrate the training activity upon which this study is based. I admire your leadership and your dedication to your staff and the students in your care.

Last but not least, my thanks go out to the trainers and teachers who participated in this study. You are all a credit to the teaching profession.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Background and Rationale

During the past three decades, performance-based tasks have been incorporated into many educational assessment programs (Bracey, 2002; Giordano, 2005; Lane & Stone, 2006; Yancey, 1999). Some experts report that such assessments were considered a step in the right direction (Linn, 2000), especially the switch from indirect, multiple-choice tests of writing skills to more direct essay-type tasks (Bracey, 2002; Cooper & Odell, 1977). While performance assessment tasks were often described as "authentic," they were criticized for their lack of reliability. Performance tasks could not be evaluated as efficiently as multiple-choice tests because "right" answers were less clearly defined (Camp, 1993; Shale, 1996). This is a particular concern in the rating or scoring of students' writings because "reliability of scores is a major necessary condition for the validity of inferences and decisions based on performance assessments" (Moore & Young, 1997, p. 4).

Test developers attempt to create efficient and reliable test instruments, but writing assessment presents a unique challenge. The challenge lies in how to create reliable assessments that reflect real-world applications (Brossel, 1986). Direct writing assessments require the test taker to respond to a given prompt or topic with a purpose and audience in mind (Bracey, 2002), and are therefore considered more "authentic" than multiple-choice tests, but direct writing assessments have also been criticized. According to Weigle (2002) the strongest and most common argument against direct writing assessment is the scoring rubric. While multiple-choice tests are characterized by their specific, "right" answers, rubrics used to score direct writing tasks are less objective. Raters who use rubrics do not necessarily use the same criteria to arrive at the same scores (Weigle, 2002; White, 1993).

Test reliability is important, but Cherry and Meyer (1993) maintained that it is inappropriate to apply the concept of reliability as understood in the sense of the classical "objective" tests to essay evaluations. They contended that essay evaluations are "multidimensional rather than one-dimensional" (p.113). Heck and Crislip (2001) similarly concluded that compared to indirect multiple-choice assessments, "direct writing assessment likely measured a more diverse set of skills" (p. 287).

Measuring progress and the products of learning are complex activities (Brennan, 2006). White (1993) explained that when evaluating student writing, the traditional concept of "true score," a term that refers to the "accurate measurement of the construct being evaluated" (p. 97), is not clearly defined because "we sometimes find differences of opinion that cannot be resolved" (p. 98). For this reason, White maintained that the best way to ensure the accuracy and fairness of writing test scores is to encourage a community of educated specialists (the raters) to adopt common goals and procedures. As assessment trends move toward applications of information and skill, well-developed rubrics with specific evaluation criteria become valuable tools for both teachers and students (Quinlan, 2006; Montgomery, 2000).

Wiggins (2006) called for continued refinement of performance-based assessments because such assessments are generally aligned to states' standards and have often led schools to strengthen instruction and curricular goals. Forty states have direct writing assessments (Ferrara & DeMauro, 2006). For example, in Florida, students in grades 4, 8, and 10 are required to demonstrate their ability to compose a piece of writing that is organized, focused, supported, and "mechanically" correct. The direct writing assessment requires students to respond to either a narrative, expository, or persuasive prompt (*Florida Department of Education FCAT Handbook,*

2005). Florida's direct on-demand writing assessment is evaluated holistically: raters score the writings as a "whole" using a 6-point rubric. Some schools and school districts in Florida recommend interim tests that follow the same format and use the same scoring rubric as the state mandated assessments and have incorporated the interim assessments into their district writing plans (Orange County Public Schools, 2008; Seminole County Public Schools, 2007). As within-school and within-district on-demand direct writing assessments become more prevalent, training teachers to reliably score those informal assessments becomes more important.

Purpose of the Study

The purpose of this study is to evaluate the effectiveness of an in-house training designed to enhance the reliability of holistic scores teachers assign to the writings of elementary school students. The general complexity of most writing tasks makes instruction and evaluation of written language a challenge at almost every grade level. If teachers are to move beyond indirect assessment using multiple choice questions, they must develop confidence in their own ability to guide students through authentic tasks. It is hoped that formal and informal training programs can help teachers better understand, and thereby refine, their roles as instructors, facilitators, and evaluators (McLaughlin & Talbert, 2006). This study also looks at teachers' evaluation of their own efficacy as writers and how that might influence their evaluation of students' writings.

Research Questions

Two questions are the focus of this study:

- Will teachers benefit from in-house professional develop conducted by staff experts designed to align teachers' evaluations of students' writings to standards established by the state?
- Will teachers' perceptions of their personal writing ability and their approach to writing tasks influence their assessment of students' writings?

Hypotheses

It is hypothesized that exposing raters to directed discussion sessions and giving them the opportunity to identify exemplars of student performance at various levels of the performance scale will significantly improve the consistency of assigned scores and align them with standards set by the state.

It is also hypothesized that teachers with low levels of writing apprehension will tend to assign lower scores to students' writings than their "less-confident" peers.

Definitions

Anchor Papers – samples of responses to on-demand writing that serve as exemplars of performance at different levels of the scoring rubric. (*Florida Department of Education FCAT Handbook*, 2005)

Direct Writing – Writing in response to a given prompt or topic with a purpose and audience in mind. (Bracey, 2002)

Expository Writing – writing that explains or gives information in an essay format.

Florida Comprehensive Assessment Test (FCAT) Writing+ - a Florida state assessment test designed to measure performance benchmarks related to writing. The fourth-grade test consists of two parts. The prompt component requires students to write a response to an assigned topic. In a second part, students answer multiple-choice questions. Students receive scores for both the multiple-choice portion and the prompt portion of FCAT Writing+. (*Florida Department of Education FCAT Handbook*, 2005)

"**High-confidence" Group** - Three participants in this study whose responses to Bowie's (1996) *Teacher/Writer Questionnaire* (Appendix A) indicated comfort and strong self-efficacy related to writing and/or writing tasks. The specific questionnaire items used to determine the "confidence" score or ranks (Appendix B) of the participants were items 2, 7, 11, 15, 20, 22, 30, 32, 34, 35, 37, and 39.

Holistic Scoring – a method of evaluating the overall quality of a piece of work.

Indirect Writing Assessment – a test of skills characterized by multiple-choice items that tap in to the test taker's knowledge of specific areas of written expression such as grammar, sentence structure, punctuation, and spelling. (Conlan, 1986)

"Low-confidence" Group – Three participants in this study whose responses to Bowie's (1996) *Teacher/Writer Questionnaire* (Appendix A) indicated discomfort or lack of self-efficacy related to writing and/or writing tasks. The specific questionnaire items used to determine the "confidence" scores or ranks (Appendix B) of the participants were items 2, 7, 11, 15, 20, 22, 30, 32, 34, 35, 37, and 39.

Narrative Writing – writing that tells a story based on a real or imagined event.

On-demand Writing - writing assessments that require students to respond to a given topic during a set time limit. (National Writing Project & Nagin, 2003)

Performance Assessment – an assessment designed to measure the degree to which the test taker can perform a task that resembles the conditions in which the skills being tested are actually applied. (Lane & Stone, 2006)

Prompt – a writing assignment that identifies a topic, a "think about it" sentence, and a sentence that gives directions. (*Florida Department of Education FCAT Handbook*, 2005)

Rater – a person who scores a piece of writing according to specific guidelines.

Rubric – a scoring tool that lists the criteria for assessing a piece of work. (Quinlan, 2006)

Self-efficacy – beliefs in one's abilities and capabilities. (Bandura, 1977)

CHAPTER TWO: REVIEW OF LITERATURE

In recent decades, there have been several shifts in the focus and style of educational assessment, especially in the area of written language. In order to understand the concepts and questions surrounding this study, following is a literature review on large-scale assessment, inter-rater reliability, and teacher attitudes and self-efficacy.

Large-Scale Assessment

History

The U.S. government has attempted to revamp and improve public education since the late 1950's (Cuban, 2004; Giordano, 2005). The 1954 U.S. Supreme Court decision in the case of *Brown vs. Board of Education*, the launch of *Sputnik* by the Soviets in 1957, and Eisenhower's National Education Act of 1958 all served to direct attention toward what was going on in public schools across the country (Cuban, 2004). However, it wasn't until Congress passed the 1965 Elementary and Secondary Education Act that standardized test results were viewed as tools for assessing the effectiveness of teachers and educational administrators (Giordano, 2005; Linn, 2000). Since then, forty-nine states have adopted standards of what their students should know and established tests to assess their performance (Cuban, 2004). Educational programs have been initiated, goals have been set, and demands for accountability have been made by both national and state legislatures. Most recently, the No Child Left Behind (NCLB) Act of 2002 put the power of the federal government behind standards-based reform and required that schools show evidence that all students are learning (Brennan, 2006). The future holds many challenges for teachers, administrators, and lawmakers. While some say that high-stakes assessments are counterproductive and not worth the monetary costs (Baines & Stanley,

2004), others like Bracey (2002) admit, "It is not likely that we will reduce our overreliance on tests any time soon" (Bracey, 2002, p. 83).

In spite of the many shortcomings associated with standardized assessments (Reeves, 2004; White, 1993), many policy makers, administrators, teachers, and parents hold firm to the belief that standardized tests measure student achievement better and more equitably than classroom teachers do (Brown, 1986; Mabry, 2004; Phelps, 2005; Reeves, 2004).

Linn (2000) outlined reasons for the appeal of the standardized assessment and attempted to explain why stakeholders and policy makers view them as a means of reform:

- First, assessment is relatively inexpensive compared to the cost of reducing class size, raising teachers' salaries to attract more able teachers, increasing instructional time, or implementing changes that require extensive training.
- Second, it is relatively easy to implement testing at the state level. Taking actions to change what happens inside classrooms are considerably more complicated.
- Third, assessment changes can be rapidly implemented within the term of office of policymakers and lawmakers.
- Fourth, the results can be used to justify the means.

Lynn further discussed the standards movement that can be linked to the Clinton Administration's Goals 2000. While adopted standards vary from state to state, Linn (2000) saw it as a small step in the right direction.

> The key point for present purposes; however, is that content standards can, and should, if they are to be more than window dressing, influence both the choice of constructs to be measured and the ways in which they are eventually measured. (Linn, 2000, p. 8)

To guard against mistakes made by individual teachers, to guard against prejudice, to guard against inequalities, policy makers, administrators, teachers, and even students guardedly embraced what many see as the objective: efficient, and "accurate" standardized assessment (Baines & Stanley, 2004; Phelps, 2005). The term "standardized" implies uniformity, and for decades test makers have attempted to improve and move beyond the age of testing when test makers were accused of "cultural insensitivity, greed, elitism, irresponsible ambition, and political ruthlessness" (Giordano, 2005, p. 191). In 1986, Lederman described America as a "test-happy" culture, and in 2006, Phelps stated, "The U.S. public has consistently favored standardized testing in the schools, preferably with consequences (or "stakes") riding on the results, ever since the first polls taken on the topic several decades ago" (p. 19).

Large-Scale Writing Assessment

According to Lunsford (1986), current attempts to assess writing dates back to 1873-74 when Harvard University administered its first written exam in English composition. Prior to that time English departments held oral examinations to test student abilities and accomplishments. Lunsford noted that the "long and acrimonious" debate between proponents of oral and written examinations ended early in the twentieth century with written exams finally claiming victory because they were viewed as more "objective" and more "fair." The College Entrance Board was founded in 1900 (Lederman, 1986), and with it came the first effort to standardize writing assessment.

Writing, defined as the communication of ideas through written language (Camp, 1993; Huot, 1993; Routman, 2000; Williamson, 1993), can be directly or indirectly assessed. Indirect

assessments are generally characterized by multiple-choice test items designed to tap into the test taker's knowledge of specific areas of written expression such as grammar, sentence structure, punctuation, and spelling (Conlan, 1986). Such tests of writing "mechanics" are easily packaged into large scale-assessments and are relatively inexpensive to administer and score. Direct assessment of writing, on the other hand, requires the students to write in response to a given prompt or topic (Breland et. al., 1987). "Standardization" is limited to the prompt itself and the amount of time students are given for the test; each response is as individual as the test taker (Bracey, 2002). Some say that direct assessments of writing are "authentic" assessments because they require the test taker to put together the various pieces of the writing puzzle with a purpose and audience in mind (Bracey, 2002; Cooper & Odell, 1977).

In recent decades, the trend in large-scale standardized testing has been to move toward more rigorous and more authentic assessments that are linked to specific content standards (Crocker, 2005; Goodman & Hambleton, 2005). While open-ended performance assessments are not new to the field of education, they are fairly new to large-scale standardized tests, (Bracey, 2002; Lane & Stone, 2006). Because of the somewhat unstructured nature of the task, performance assessments garner more criticisms related to reliability. In essence, researchers report that performance assessments present opportunities for more mistakes to be made (Camp, 1993; Shale, 1996). Multiple-choice tests generally pass reliability inspections with flying colors. Camp states that, "although human judgment is involved in the development and selection of test items, the multiple-choice test is unaffected by the subjectivity of human scoring" (Camp, 1993, p. 47). She sums up the quandary:

> Multiple-choice tests offer reliability, efficiency, economy, some contributions to validity, and a convenient basis for statistical comparisons from one test to another, even

> though they measure only a limited number of subskills for writing, and measure them only indirectly. The impromptu writing samples provide a demonstration of the writer's handling of both the subskills for writing and the larger-order skills involved in actually composing text: generating and developing ideas, organizing, establishing connections within the text, and finding the tone and rhetorical stance appropriate to the topic and audience. (Camp, 1993, p. 51)

From the 1950s through the 1970s, when the results of large-scale assessments drew attention to the shortcoming of the American public education system (Gredler, 2001), it was the norm to assess writing through "objective" multiple-choice tests (Bracey, 2002; Yancey, 1999). Multiple-choice items dominated large-scale assessments because ease of administration and objectivity were perceived strengths (Bracey, 2002; Yancey, 1999). Measurement of writing ability was limited to grammar and mechanical structures (Salies, 1998).

In the 1970s concerns about test validity grew. As attitudes about the goals of education changed (Yancey, 1999), criticisms of the multiple-choice standardized assessment surfaced. One argument was that multiple-choice items solicited superficial rather than thoughtful responses (Lyman, 1998).

A test is considered valid if it "prompts students to represent the dimensions of the learning desired" (Maki, 2004, p. 93). In other words, if the purpose of writing assessment is to determine how well a student can write, it seemed logical that the test should involve writing (Bracey, 2002; Diederich, 1974), not bubbling in answers on an answer sheet. The testing of a learned skill through more direct applications that require students to construct or create responses have gained popularity in recent decades in spite of reliability issues that naturally

evolve due to the subjectivity associated with scores being assigned by human raters (White, 1993). Even though the assessment of direct writing may be fraught with inconsistencies, Salies (1998) stressed that direct measures of writing are the "right" thing to do. She stated that research studies

> support the contentions that direct measures tap a production factor, and thus represent a separate construct from that of indirect tests namely, the ability to write as opposed to knowledge of conventions of writing. Indeed, nothing seems more logical than requiring students to actually write to gauge if they can do it. (pp. 4-5)

Conflicts between Theory and Practice

Cognitive theorists such as Jean Piaget (1965) described learning as complex and learner-specific. Piaget's experiments, characterized by the direct observation of children, led him to conclude that learning occurred through the processes of assimilation and accommodation. Educators who embrace Piaget's theory view children as active learners who construct knowledge through the interaction of what they already know and their involvement with content through activities that promote active engagement, inquiry, problem solving, and collaboration (Pilcher, 2001). In stark contrast to the cognitive learning theories, radical behaviorism as presented by B. F. Skinner focuses on environmental conditions and observable responses (Gredler, 2001). It is argued that behaviorist theory oversimplifies the leaning process, and as theories go, it comes up short (Driscoll, 2000). Its influence however has taken firm root as is evidenced by the educational community's love affair with standardization and "objective" measures (Mabry, 1999; Neill, 2000; Phelps, 2005). While few of today's educators hold to radical behaviorist pedagogy, there is a definite conflict between what teachers believe about

learning and what standardized tests measure (Mabry, 2004). “In standardized testing, the manifest theory is *behaviorism*, the idea that changed behavior, as an educational product, can be measured” (Mabry, 2004, p 116).

In contrast to behaviorists, cognitive scientists, influenced by the cognitive stage theory of Jean Piaget (1965), see the child as an active learner constantly interacting with the environment. Theorists such as Jerome Bruner (1971) and Lev Vogotsky (as cited in Rieber, 1997) believed the isolated and mechanical presentation of rules should be avoided, not only because such methods hinder self-expression, but also because it strips meaning from the learning process and thereby limits the learner’s development. While few schools demonstrate pure, child-centered instructional methods, in recent decades there has been a slow but steady movement away from the teaching of isolated skills and toward a more humanistic approach to instruction (Ornstein & Hunkins, 1998).

Vygotsky’s theory of cognitive development has strongly impacted educational practices (Gredler, 2001; Langford, 2005). The theory holds that both biological and cultural factors contribute to cognitive development. One of the main tenants as it relates to writing is that instruction should be organized in such a way to engage the learner in meaningful activities (Gredler, 2001). Some individuals in the educational community consider the movement away from the structured multiple-choice assessments a step in the right direction (Wiggins, 2006).

Writing Assessment Trends (1950 – Present)

Yancey (1999) separated the recent history of writing assessment at the post-secondary level into three overlapping waves. The first wave was roughly a 20-year period that began around 1950 and continued until 1970. During that time period, the objective multiple-choice test

dominated large-scale assessments. The second wave, roughly set in time from 1970-1986, was the period when the holistically scored essay gained prominence. Around 1986, the third and current wave built on and expanded the second. The "one essay" model was replaced by writing collections known as portfolios. As trends go, it can be noted that the first two waves Yancey identified at the university level are mirrored by large-scale assessments at the elementary level. While classroom assessments were likely to contain multiple work samples, commentaries, and reflections, large-scale portfolio assessments were deemed impractical and not cost efficient.

Currently, large-scale writing assessments are in a tug of war, with validity and reliability seemingly on opposite ends of the rope. The dilemma is that stakeholders don't want one without the other. As far back as 1986, Conlan tackled the assumptions related to multiple-choice and essay type tests and concluded that it might be best if test developers used the advantages of multiple-choice questions and the advantages of essay questions by incorporating both methods for assessing writing. Her argument was that no test is completely "objective," and that the term objective has been erroneously equated with the fact that the test can be scored by a machine or by people who make no judgment about predetermined answers. She argued:

> The test itself is not objective; it does not function as an ideal criterion against which all and everything in a subject can be measured. It fulfills a certain purpose and only that purpose. Most important, it was designed and put together by human beings who, like all other human beings, have faults and have opinions. Those opinions inform what is measured by the test they have developed and how what is measured is measured... When one considers that all testing is merely a matter of sampling a particular universe of skills or knowledge, any emphasis placed on a particular facet of writing - either by the number of questions assigned to the measurement of a particular skill or bit of knowledge

> or by the kinds of questions used and the types of writing problems tested – helps the testmaker made a clear statement about writing. (Conlan, 1986, p. 110)

Conlan (1986) contended that it is narrow-minded to assume that an essay test is a better measure than a multiple-choice test. It is generally accepted that multiple-choice writing assessments lack face validity, or the ability of the testing instrument to measure what it intends to measure (Lyman, 1998), but an essay is not always better. Each essay prompt presents a unique writing problem, the knowledge a student brings to the topic is unique, and the rater or scorer carries a unique set of "experience baggage" as well. Such differences create reliability problems, and there is little value in content validity if there is little or no reliability (Best & Kahn, 2003). According to Conlan (1986), compromise may be the best way to solve the dilemma. The test developer can use the best of both worlds, the stronger reliability of the multiple-choice assessment and the validity of the essay. Florida is one state that has moved toward just that type of combined writing assessment. In 2005 when it changed from the "essay only" FCAT Writes! to a combined format now formally called FCAT Writing + (*Florida Department of Education Keys to FCAT*, 2007).

Writing Rubrics

Two types of rubrics are commonly used for direct writing assessment, holistic rubrics and analytic rubrics. Holistic scoring, which is based on an overall impression of quality, dates back to the 1970s (White, 1993; Yancey, 1999). Raters using holistic rubrics do not quantify the particular strengths and weaknesses of a piece or writing, but consider to what extent and in what manner the writer accomplishes the overall goals such as purpose, organization, style, and

conventions (Baldwin, 2004; Cooper & Odell, 1977). In contrast, raters using analytic rubrics assign scores to individual writing traits that are then tallied to determine an overall score (Diederich, 1974). Not surprisingly, there are advantages and disadvantages associated with the use of both types of rubrics.

Bainer & Porter (1992) conducted a study that identified concerns of five third-grade teachers as they attempted to implement holistic scoring procedures in their classrooms. Prior to the study, the teachers at the school received training on using the rubric during two, one-hour training sessions. The first session introduced the district-approved rubric, and the second session allowed for scoring practice. After the initial training, the study focused on five participants' reactions while evaluating writing samples of 30 students equally distributed among their 5 classrooms. There were an equal number of samples from high, middle, and low ability students. During the yearlong study, 187 statements of concerns were collected and categorized as teachers recorded their thinking while deciding on the rubric scores. It was found that 50.3% of the "concern" statements were associated with the general usability and interpretation of the rubric. While the rubric was generally praised for its efficiency, teachers reported being troubled by "vague" wording at the various levels of the rubric. Bainer & Porter reported:

> Because of their analytical mindset, some teachers related "agonizing" experiences associated with assigning a rubric score. Others reportedly vacillated between adjacent rubric scores. Some teachers considered the holistic approach a matter of personal judgment, the rating scores thus showing differences in teacher expectations. (p.19)

Waltman, Kahn, & Koency (1998) compared Analytic-Impression and Focused-Holistic methods, both which used a single score report format. For purposes of that study, the Focused-

Holistic format was defined as a scoring method that collectively summarized criteria and did not "emphasize the dimensional aspect of each criterion" (p. 3). Conversely, the Analytic-Impression method separated the criteria for each score point, but unlike true analytic scoring, raters determine which "overall score 'best fits' the student's performance" (p. 3). The study participants evaluated students' responses to middle school performance tasks for science. It was determined that the Focused-Holistic method had higher inter-rater agreement rates than the Analytic-Impression method. Scorers using the Focused-Holistic method were in exact agreement 60% of the time, and agreement was within one score point 93% of the time. Scorers using the Analytic-Impression method were in exact agreement 49% of the time, and agreement was within one score point 88% of the time. The authors reported that study participants preferred the Analytic-Impression method for obtaining diagnostic information, but when asked which scoring method was easier to use, raters were divided. The researchers speculated that the efficiency of the Focused-Holistic method won out over the more labor-intensive Analytic-Impression method.

Advantages and Disadvantages of Holistic Scoring

Weigle (2002) compared holistic scoring procedures to analytic scoring procedures and reported that the analytic scoring rubrics, designed to evaluate several aspects of writing criteria, provided more detailed information about the writer's performance. The main disadvantage of the analytic rubric was the increased time that analytic scoring demanded. Similarly, Bainer and Porter (1992) found that "teachers are more apt to engage their students in writing if it doesn't take a lot of time to grade the papers" (p. 11).

For state-mandated, large-scale assessments, efficiency is usually a major consideration (Breland et. al., 1987). Florida is one state that utilizes a holistic scoring rubric to evaluate students' writing performance (*Florida Department of Education FCAT Handbook*, 2005). The Florida Comprehensive Assessment Test (FCAT) Writing + is a performance-based assessment administered to students in the fourth, eighth, and tenth grades throughout the state of Florida. One part of the FCAT Writing+ is a direct writing task. For that part of the assessment, raters use a holistic scoring method designed to evaluate how well a student integrates four elements of effective writing: focus, organization, support, and conventions. A six-point holistic rubric is used to evaluate performance levels (*Florida Department of Education FCAT Handbook*, 2005).

Defending the use of holistic rubrics, Cooper and Odell (1977) argued that a piece of writing has a purpose and communicates a whole message; therefore, holistic assessments bring us closer to what is important in written communication. Camp (1993) took a similar position:

> In many respects, the holistically scored writing sample fares better than the multiple-choice test with respect to validity. As a performance measure drawing on the broader range of skills and strategies necessary for actually generating a piece of writing, the writing sample has frequently been seen as a more valid form of assessment. It allows students to demonstrate skills not tapped by the multiple-choice test and more compatible with the current theoretical construct for writing and with desirable practice in writing instruction. (p. 49)

Breland et. al. (1987) stated:

> The chief assumption that underlies holistic scoring of essays is that the whole text or composition is more than the sum of its parts. To look at a composition from the aspect of

> its mechanics, its rhetorical structure, its syntactic patterns or complexity, or its handwriting is to view it narrowly. To look at a composition as a whole in order to judge its quality as an entity itself is to score it holistically. When the number of compositions to be scored is large, holistic scoring is the most practical method. For that reason it is most often used in large-scale assessments. (p. 18)

Lederman (1986) argued that since testing is so well entrenched in our social and educational system, and since writing is a critical skill for success in our country, it is important that we choose a type of testing instrument that supports what we value. Lederman (1986) states: "Yet testing, which should be an outgrowth of and subordinate to curriculum, in reality often drives curriculum" (p.41). It might be considerably easier to face and come to terms with a test-driven curriculum if the test that drives it is in line with what it valued. Wiggins (2006) states:

> Practical alternatives and sound arguments now exist to make testing once again serve teaching and learning. Ironically, we should "teach to the test." The catch is to design and then teach to standard-setting tests so that practicing for and taking the tests actually enhances rather than impede education. (p. 252)

Salies (1998) expresses a similar position when she stated: "If teaching to the test occurs, it is far more desirable to have teachers training students to pass a writing sample than an objective test" (p. 5).

Performance indicators outlined in rubrics can identify a student's strengths as well as weaknesses (Maki, 2004). "Ideally, interpreting patterns of weakness leads to adjustments or modifications in pedagogy; curricular, co-curricular, and instructional design; and educational

practices and opportunities" (Maki, 2004, p. 121). Higgins, Miller, and Wegmann (2007) pointed out that sound instructional approaches help children develop into competent writers, and as students' skills develop in response to good teaching, "teaching to the test" becomes unnecessary. Higgins, Miller, & Wegmann stated:

> After all, assessment is a component of instruction and not an end unto itself. Assessment should help the teacher learn about individual strengths and needs of students for purposes of instruction. The goal of instruction is to produce lifelong learners, not test takers. (Higgins, Miller, & Wegmann, 2007, p. 311)

Mabry (1999) concluded that while the rubrics used in direct assessment of student writing are a cut above the traditional multiple-choice test items, they tended to standardize writing instruction. More recently, Higgins, Miller, and Wegmann (2007) argued that good teaching practices need not be abandoned in an effort to raise test scores. After conducting a survey of states' standards and tests, Higgins, Miller and Wegmann concluded that the goals and practices promoted through the use of the writing process (planning, drafting, revising, and editing) and the analytic scoring system known as 6 + 1 Traits (Northwest Regional Educational Laboratory, 2004) were to some extent reflected in the standards of all 50 states.

Inter-Rater Reliability

Measuring Reliability

While many have argued in support of direct assessment of writing as opposed to indirect, multiple-choice assessments (Camp, 1993; White, 1993; Wiggins, 2006), the use of rubrics and the human scorer necessary to the direct assessment process naturally generated

concerns of test reliability (Shale, 1996; White; 1993). Inter-rater, or inter-scorer, reliability is the degree to which consistency in judgments exits among raters or scorers of essays (Best and Kahn, 2003; Hopkins, 1998). "Undoubtedly, the major factor responsible for the complexity of the concept of reliability in the context of essay testing is the subjective scoring process" (Shale, 1996, p. 77).

Many approaches to determine inter-rater reliability have been discussed and promoted by researchers. Correlation approaches, percentage agreements, and generalizability theory all attempt to evaluate how consistently raters rate essays (Moore & Young, 1997; Shale, 1996).

Cherry and Meyer (1993) described three assessment scenarios and the appropriate statistical approach for each.

- Situation 1: When a large number of raters rate a portion of the writing samples, a one-way random effects analysis of variance model was recommended.
- Situation 2: When all raters rate the same essays and the ratings are described as objective, a two-way random effects analysis of variance model was considered appropriate.
- Situation 3: When all raters rate the same essays but the ratings are considered relative to other samples in the group, Cronbach's alpha is the recommended statistic because it treats the raters and the writing samples as independent variables.

While this study resembles the one Cherry and Meyer described in Situation 2, other researchers recommend intraclass correlations such as Cronbach's Alpha for reports of research (Atkinson & Murray, 1987; Moore & Young, 1997), so this study was framed accordingly.

Moore and Young (1997) presented the pros and cons of the most commonly used methods for assessing inter-rater reliability. They pointed out that correlations are appropriate when raters are evaluated in pairs and when the range of categories is broad. However, in most of today's performance assessments, the range of categories is quite small, as in the 0 to 6 point FCAT Writing + rubric. Percentage agreement, on the other hand, is more "straightforward;" it is a simple comparison of the number of times scorers or raters agree. As in the correlation approach, this method is criticized when used with scales of only a few points because "the likelihood of raters agreeing by chance alone increases, and the percentage agreement is inflated accordingly" (Moore & Young, 1997, p. 7). While Moore and Young favor the use of generalizability theory, it is a complex method more appropriate for large-scale assessments. According to Wint-Tat Chiu, (2001) generalizability theory has two major functions: "1) to evaluate the quality of the measurement procedures; and 2) to make projections about how to improve the quality of the measurement procedures" (p. 1). In the words of Reckase (1997), "As the inferential leap needed to interpret the score increases, the psychometric support for that leap increases" (p.11).

This present study used writing samples that had been previously scored by trained raters for the Florida Comprehensive Assessment Test (FCAT) Writing +. Cronbach's Alpha was used to assess inter-rater reliability, and one- sample *t*-tests were used to compare the group means to the scores assigned by the FCAT raters. Percentage agreement analyses were also used to compare the school-based scores to the FLDOE scores.

The Holistic Writing Rubric: Reliability Issues

According to Fisher, Brooks, & Lewis (2002) the "fitness for purpose requirement" should be central to all test development work. A valid test is one that replicates, as closely as possible, real world applications. Unfortunately, ensuring validity and reliability is not an easy task. The problem is that while direct writing assessments are considered valid, they are also less objective, and thereby more prone to reliability issues (Fisher, Brooks, & Lewis, 2002).

Atkinson and Murray (1987) noted that when human raters are involved in the assessment process, they become the measurement instrument. "The measurement instrument is really the rater, a person sorting written products according to the categories assigned by the researcher. Therefore, the issue of reliability is bound up in many factors" (Atkinson & Murray, 1987, p.13). Atkinson and Murray concluded that inter-rater reliability could be improved by clearly defined scoring categories.

Compared to the reliability assessments of the "objective" multiple-choice test, reliability as it relates to the human rater is much more complicated and "messy." Mabry (1999) contended that rubrics promoted inter-rater reliability by limiting the possible scores. Commonly used holistic writing scales have scores that range from 1-6 thereby giving raters a better chance of agreement. Limited scales combined with the need for raters to strive for agreement can combine to create a consistency among scorers that "reflect collective tunnel vision rather than informed consensus about the quality of student writing" (Mabry, 1999, p.4).

In Mabry's (1999) words: "Perhaps rubrics could be devised that have the comprehensiveness and flexibility to accommodate different genres, voices, and styles of writing. But perhaps writing is too personal and varied an experience to be amenable to scoring rubrics" (p. 9).

Mabry's position was supported by a study conducted by Fitzpatrick et. al. (1998). In that study, 115 raters re-scored reading, writing, language, mathematics, social studies and science performance assessment tests administered to third, fifth, and eight grade students as part of the Maryland School Performance Program (MSPSP). The researchers reported that Pearson correlations and percentage agreement for the scores assigned in 1991 and those assigned in 1992 were high for all tested subjects except writing. More specifically, an alpha coefficient of .85 or higher was reported for all subject areas except writing. The coefficient alphas for writing were in the .50s for third and fifth grade writing tasks and in the .60s for the eight grade tasks. Fitzpatrick et. al. concluded that, "raters are likely to be more consistent when they are using scoring rules that refer to observable qualities in students' responses than when they are using rules requiring that abstract qualities be inferred from a student' s response" (Fitzpatrick et. al, 1998, p. 207). Similarly, in a study of rater agreement on IQ and achievement tests, Van Noord and Prevatt (2002) found that the writing samples contained the highest number of errors and that the scoring errors were made across all scorer experience levels: novice, intermediate, and advanced.

When Cabrillo College in California conducted an investigation to validate their holistic English assessment, a total of 3,932 essays were reviewed for inter-rater reliability (Willett, 2001). Essays readers received training on the scoring rubric and they were also provided examples of writing at the different levels. The essay evaluations were used to determine placement. In surveys secured from students, 75% indicated that they felt they had been accurately placed. Ninety percent of the instructors reported that students were able to pass their classes. It was noted that the assessment process worked best when identifying candidates for the highest-level classes, and that by law, enrollment in basic skills classes could not be restricted. In

sum, the researchers reported an inter-rater reliability correlation of 0.83, with the initial two ratings being in exact agreement 81.7% of the time. The study validated the use of the holistic scoring method that had been in place at the community college since 1994.

Cherry and Meyer (1993) examined some of the issues associated with inter-rater reliability and called for a standardization of statistical methods used to report reliability of holistic scores. Cherry and Meyer (1993) stated: "A wide variety of coefficients have been reported in the compositions research and testing literature, with no discussion of, or apparent consensus on, which statistics are appropriate for which circumstances" (p. 116). While various intra-class correlation formulas exit and are utilized by researchers (Cherry & Meyer, 1993), there is general agreement that high "interrater reliability coefficients in the 80s and 90s can be reached with careful training and monitoring of raters" (p. 135).

According to the Florida Department of Education's (FLDOE) *FCAT Handbook* (2005), scorers of the FCAT Writing test are carefully trained and monitored. They attend multiple-day training sessions and are provided multiple opportunities to practice scoring. At the end of training, they must pass a qualifying exam. In addition to the scoring rubric that described the work demonstrative of each performance level, anchor papers illustrating each level of performance are readily available during the scoring process. In fact, inter-rater reliability reports are used to ensure consistency and reliability of scoring. The Florida Department of Education *FCAT Handbook* (2005) states:

> Each scorer's (or rater's) score of a student response is compared to the other score given to that response. A cumulative percent of agreement between the two scores on every response (as opposed to validity responses only) is reported for each score as the inter-rater reliability percent. The information on this report indicates whether a scorer is

> agreeing with other scorers scoring the same responses. Analysis of the report is used to determine if a scorer or group of scorers is drifting from the established guidelines and require additional training. (p. 73)

The FCAT Writing+ essays and narratives are hand-scored by at least two trained raters. If the two raters' scores are within one point of each other, the essay is assigned a score that is the average of the two scores. If the scores differ by more than two points, a third rater scores the essay. If that raters' score agrees with either of the previous scores the final score is the matched score. If the third rater's score differs from the first and second rater's score by one only point, the final score is the average of the two adjacent scores (*Florida Department of Education FCAT Handbook,* 2005).

Exact agreement when using holistic scoring is not easy to achieve. On the 2007 FCAT Writing +, 41% of the scores assigned to fourth grade essays and stories were an average of two scores. That pattern is not unlike other years. In 2006, 38% of the assigned scores were an average of two scores, and in 2005, the percentage was 40% (Florida Department of Education FCAT Writing Scores [Data file], 2007).

Support for Florida's methods for score resolutions can be found in the writing of White (1993), a pioneer in the use of holistic scoring of student essays, who attested that some writings resist agreement, and despite scoring guides and sample papers raters do not always focus on the same aspects of the writing. White (1993) suggested that the paper is simply an average or combination of the two scores and those differences of more than one point should be resolved by a third rater. He wrote:

> In fact, historically, such differences about value in most areas of experience tend to be more valuable than absolute agreement; they combine to bring us nearer to accurate evaluation than would simple agreement (that is, my score is a bit low, probably, and yours is a bit high). This is the same principle that allows us to judge work from the past in light of much critical discussion and compromise. The same is true in measurement of writing ability, where some disagreement (within limits) should not be called error, since, as with the arts, we do not really have a true score, even in theory. (p. 98)

White (1993) cautioned that since reliability is often an "underlying theoretical and practical problem" for holistic essay evaluations, the educational community needs to refrain from using them as infallible measures, but should instead consider such scores as reliable single measures in a collection of multiple measures.

The Challenge of Writing Assessment

Assessing writing is a formidable task (Bainer & Porter, 1992). Many researchers considered scoring guides or rubrics a means of bringing a level of objectivity to a subjective task, but when using holistic scoring, reliability, or consistency of the measure, is a major concern. Teachers using the same rubric have been known to arrive at different scores (Breland et. al., 1987; Shale, 1996). Bainer and Porter (1992) maintain that holistic assessment holds promise, but it has not yet been determined that teachers can consistently evaluate students' writing.

In a study conducted in Montgomery County, Maryland, Myerberg (1996) evaluated the strength of inter-rater consistency on math and language arts assessments scored by that school

district's staff. Teachers received extensive training prior to the scoring sessions and quality controls were conducted throughout the process. During training workshops, teachers discussed the scoring process and reviewed scoring rubrics. Test papers previously scored by an "expert" group, were re-evaluated by the trainees who openly discussed their reasoning and justification for assigning scores. The actual tests were scored in a centralized setting with random monitoring to ensure quality. Correlations between scorers and the percent of differences greater than one point were used to evaluate scoring consistency. Myerberg (1996) found that the task of ensuring inter-rater reliability was much more difficult for language arts assessments and that active monitoring was required to achieve consistent scoring. Objectivity in educational assessment is easier to achieve when there is one or a limited number of "correct" responses. For this reason, "proponents of objective assessment were able to accommodate the content in some academic subjects more easily than that in others" (Giordano, 2005, p. 31).

One can understand why objectivity is more easily accomplished when dealing with mathematics performance assessments because there are limits to how a student can accurately respond to particular mathematics problems. However, such is not the case in writing. Writing mechanics aside, student responses to writing prompts vary greatly. That variability combined with the descriptive nature of the holistic writing rubric has caused confusion and concern. The argument that it is difficult to separate the elements of good writing into independents factors is valid, so the challenge of providing fair and accurate assessment remains an issue in many schools and districts. Penny, Johnson, and Gordon (2000) suggested that one possible way to improve inter-rater reliability is to allow raters to augment their scores. In that study, raters used a six point holistic scale to rate fifth-grade essays, but were allowed to augment scores with a plus (+) or minus (-). Both novice raters and experts chose to augment close to 50% of the papers

they evaluated, and inter-rater reliability improved significantly. While the study did not attempt to investigate the motivation behind the augmentation, these results accentuated how difficult it was for raters to identify benchmark proficiency.

Augmentation (Penny, Johnson, & Gordon, 2000) may not be feasible or desirable in some instances, but calibration to ensure that raters respond consistently can be developed over time with successive applications. Maki (2004) outlined a training process to establish inter-rater reliability that involves several steps:

- independent scoring
- discussion among raters to review responses
- discussion to reconcile differences
- repeating the process of independent scoring
- reviewing responses again
- discussion to reconcile differences

Maki (2004) pointed out that this process repeated until raters reach consensus ordinarily takes two or three sessions. Moore & Young (1997) stated:

> The good news from the measurement literature related to performance assessment is that high rater reliability is quite possible and feasible with as few as two, and even one rater, if there are specific scoring guidelines and sufficient training for the raters. (p. 11)

Changing Goals and Strategies

It is Wiggins' (2006) position that, if done right, testing can serve to complement and enhance teaching and learning. In his opinion, there is nothing wrong with "teaching to the test,"

if the test has integrity: "genuineness, effectiveness, and aptness of the challenge" (p. 261). According to Wiggins, tests of performance tasks should be developed first; scoring and reliability should be secondary. Academic tests should be "standard-setting" tests not just standardized. Wiggins stated:

> Reform of testing depends, however, on teachers' recognizing that standardized testing evolved and proliferated because the school transcript became untrustworthy. An "A" in "English" means only that some adult thought the student's work was excellent. Compared to what or whom? As determined by what criteria? In reference to what specific subject matter? The high school diploma, by remaining tied to no standard other than credit accrual and seat time, provides no useful information about what they have studied or what they can actually do with what they have studied. (p. 253)

Higgins, Miller, and Wegmann (2007) reported that most states have writing tests that requires students to write in response to a prompt, and they noted that, "If students are given daily opportunities to write meaningful texts while learning the different genres of writing, they will develop fluency and be able to write in that genre if asked to do so on a writing test" (p. 316). Lumley and Yan (2001) found that the Pennsylvania state writing assessment prompted teachers to provide students with varied and frequent writing opportunities. Lumley and Yan (2001) reported:

> The key factor influencing classroom practices is teacher training. Training significantly affects the value which teachers place on writing assessments and how frequently they utilize teaching strategies connected directly to the skills and demands of the state writing assessment… those teachers who have been trained in holistic scoring more strongly

agree that the state writing assessment has improved the writing ability and skills of their students as well as their ability to teach writing. (pp. 33-34)

Teacher Attitudes and Perceptions of Self-Efficacy

Teachers as Assessors

In recent years, the power of tests to shape the curriculum has increased tremendously (Mabry, 2004). Abrams, Pedulla, and Madaus (2003), reported on a national survey conducted by the National Board on Educational Testing and Public Policy. They found that 70% of teachers from high-stakes environments reported that they were preparing students for the state test throughout the school year. Mabry (2004) noted that while multiple measures of student progress and achievement still existed in most classrooms, the reality was that in states with high-stakes assessment programs, school failure was often based on a single set of test scores (Mabry, 2004).

The need for teachers to be "effective agents of assessment" is reiterated in the *Standards for the Assessment of Reading and Writing* (International Reading Association and National Council of Teachers of English Joint Task Force on Assessment, 1994). The authors stressed the socially complex nature of reading and writing, and acknowledged the challenges associated with making assessments fair and equitable.

> Teachers cannot be expected to acquire and refine this knowledge without considerable support. Indeed, the major investment required for improving assessment must be in staff development and school-community learning. Serious attention must be given to providing the time and conditions that will help teachers maximize and reflect on their

knowledge. (International Reading Association and National Council of Teachers of English Joint Task Force on Assessment, 1994, p. 29)

According to Huot (2002), "Assessing, testing, or grading student writing is often framed as the worst aspect of the job of teaching student writers" (p. 63). The negative feelings associated with assigning grades to students' written work may be linked to teachers' personal experiences with writing and their perceptions of their writing ability. In one study, a heterogeneous group of 226 student teachers preparing to teach in elementary and secondary schools responded to a survey designed to measure their writing apprehensions and beliefs about their future role as teachers (Bowie, 1996). Bowie (1996) found that respondents having low levels of writing apprehension were much more confident about their ability to evaluate and critique another person's writing. While 86% of respondents to Bowie's survey reported a belief in the importance of writing across the curriculum, only 40% felt confident about evaluating the writing of others, and 45% felt that writing assignments were difficult to grade.

The Standards for the Assessment of Reading and Writing (International Reading Association and National Council of Teachers of English Joint Task Force on Assessment, 1994) stated:

> Assessment instruments have traditionally been conceived of as tests and teachers have been viewed as merely consumers of information generated by these tools. However, of all the evaluation that takes place in education, most take place in the classroom, as teachers and students interact with each other. Teachers design, assign, observe, collaborate in, and interpret the work of students in their classrooms. They assign meaning to interactions and evaluate the information that they receive and create in these

settings; in short, they do function as agents of assessment, and their assessments have enormous impact on students' lives. (p. 27)

Teachers' Self-Efficacy

Teacher attitudes about standardized tests are often determined by the impact such tests have on their own perceptions of their efficacy (Salpeter & Foster, 2000). It goes without saying that "low" test scores can be demoralizing to both teachers and students, but "high" test scores can have the opposite affect. Some school districts have welcomed the high-stakes tests as validation of what they promote in their schools and individual classrooms.

Social-cognitive theorist Alfred Bandura (1977) defined self-efficacy as "people's beliefs in their capabilities to mobilize the motivation, cognitive resources, and courses of action needed to exercise control over environmental demands" (p. 191). Zimmerman & Schunk (2003) pointed out that self-efficacy should not be equated with self-concept. Self-efficacy involves self-appraisal and refers to the belief that one can perform particular academic tasks.

Bandura (1986) identified four things that influence an individual's beliefs about their capabilities. They included mastery experiences, vicarious experiences, social experiences, and emotional states.

- Mastery experiences were defined as personal successes an individual has with a particular task. Mastery experiences are especially important because they provide evidence of capability and shape an individual's reaction to future tasks that are similar. Mastery experiences often determine whether an individual expresses an "I can do" or "I can't do" attitude.

- Vicarious experiences were defined as a person's observations of others experiencing success with a task. Vicarious experiences affect the learner when he or she compares himself to people he or she observes.
- Social experiences were characterized as verbal persuasions. They are described as instances when others attempt to persuade the individual that he or she is capable of attaining success with a particular task.
- Emotional experiences were classified as physiological responses or feelings associated with a particular task or behavior. Emotional experiences, or internal arousal states, are sometimes debilitating, but can just as frequently lead to increased effort.

However the success-failure experiences played out, it was noted that success rates naturally rose as an individual's concept of self-efficacy rose, and one or two failures did not necessarily tip the scale in the "low" direction. If there was enough positive mastery, vicarious, and social experiences related to the task on hand, the individual was better able to view failure as a minor setback.

In the words of Tschannen-Moran and Hoy (2007): "Teachers' self-efficacy is a little idea with a big impact. Teachers' judgments of their capability to impact student outcomes have been consistently related to teacher behavior, student attitudes, and student achievement" (p. 54). In their study, Tschannen-Moran and Hoy (2007) found that mastery experiences had the greatest impact on self-efficacy judgments of both novice and experienced teachers. While novice teachers reported lower levels of self-efficacy than their more experienced peers, verbal persuasion in the form of support of parents, colleagues, administrators, and the community was also reportedly more important for novice teachers than experienced teachers.

Teachers' Self-Efficacy and the Impact on Instruction

Graham et. al. (2001) tested the construct validity of a self-reporting instrument designed to measure whether primary grade teachers' efficacy in writing was related to their beliefs about how to teach writing. They found that the new instrument was valid and that "high-efficacy" teachers reported that their students spent more time engaged in writing tasks. "High-efficacy teaches also reported that they spent more time teaching the mechanical aspects of writing along with the activities associated with the writing process. While their instruction included lessons designed to teach grammar and usage, the "high-efficacy" teachers reported that they emphasized the natural and authentic activities. In sum, the participants' feelings of efficacy were related to the amount of time allotted to writing activities and the type of instructional activities and methods they utilized.

In a study conducted by Hoy and Spero (2005), changes in teacher efficacy were examined through the use of multiple self-reporting instruments. Expanding on Bandura's theories, Hoy and Spero (2005) examined how novice teachers' sense of efficacy changed during pre-service training and during the first year of teaching. Three instruments were used in the study: Bandura's assessment of Instructional Efficacy, Gibson and Dembo's Teacher Efficacy scale (as cited in Hoy & Spero, 2005), and an instrument designed to measure general teaching efficacy (GTE) and personal teaching efficacy (PTE). In the Hoy and Spero study, the participants were 53 prospective teachers enrolled in a Master's of Education initial teaching certification program. Changes during the course of pre-service and during the first year of teaching were noted. For all four instruments used in the study, self-efficacy increased from the beginning to the end of the teacher education program. Significant decreases in measures of self-

efficacy were noted on the Bandura and GTE scales at the end of the first year of teaching. The researchers guardedly suggested that the optimism and self-confidence that develops during pre-service teaching programs might simply come down to earth and reality during the first year of teaching. Tschannen-Moran & Hoy (2007) suggested that the drop in self efficacy during the first year of teaching could be problematic since teachers' beliefs, particularly beliefs about their own self-efficacy, are related to the effort they put into teaching (Tschannen-Moran & Hoy, 2007).

Teachers' Attitudes about Standardized Tests

Today's standardized tests are considerably better constructed and much fairer than they were back in the latter half of the twentieth century (Reckase, 1997; Lyman, 1998), but these highly valued measurement tools will never be flawless (White, 1993; Lyman, 1998). There is continuing debate over whether the current high-stakes testing improves instruction (Abrams, Pedulla, Madaus, 2003). On that subject, Winkler (2002) found differences in attitudes of novice and veteran teachers. In intensive interviews with a small group of Virginia schoolteachers over the course of a school year, Winkler found that teachers' perceptions of the value of standardized testing was directly related to their time on the job. While experienced teachers expressed frustration with "prepackaged" curricula, the novice teachers' views were quite different. Novices reported that uniformity of curriculum promoted by Virginia's Standards of Learning (SOL) test was a way to maintain equity and ensure collaboration. Novice teachers reported that the SOL test gave them direction and purpose, while experienced teachers complained of loss of power and personal choice.

Brindley and Schneider (2002) surveyed fourth grade teachers from a large school district in the Southeastern United States in an attempt to gain insights into the teachers' instructional practices as well as their perceptions of writing development in young children. The teachers in the study worked in a high-stakes testing environment evidenced by the fact that students in the district must receive a score of 3 or higher on the 6 point writing assessment scale in order to be promoted to middle school. While teachers reported being limited in their use of instructional approaches and strategies, they also admitted that the writing test preparations brought more structure, organization, and formality to their instruction. One dominant test prep strategy reported by the teachers in the survey was the district-prescribed method of requiring students to respond to narrative and expository topics or prompts. This type of on-demand writing within a given time limit was viewed as essential to ensuring test success. Teachers reported that they modeled writing more often, and they had actually "moved beyond the basics" (p.334). The study showed that teachers had adapted their practice as a result of assessment pressures, and that both they and their students felt substantial stress related to the high-stakes assessment process.

Attitudes Linked to Training and Experience

In most states, elementary teachers receive very limited training on how to teach writing (Routman, 2000). Knowing little about how to approach writing instruction, many teachers develop unrealistic expectations about the capabilities of their young students, often focusing on grammar, punctuation, spelling, and capitalization because standards for those elements of the writing process are well known. Routman (2000) insisted that specific grade-level writing rubrics should establish a picture of good writing that includes organization, the use of engaging

language, and voice. With a comprehensive rubric and exemplar papers to guide them, teachers can set goals for themselves and their students.

Some teachers mentioned a belief that they must themselves be writers in order to effectively teach writing (Cremin, 2006). “If teachers engage as writers, taking part in the creative process of composing, they arguably will be in a stronger position to develop the creative voice of the child” (p. 418). Cremin (2006) attempted to analyze teachers as they developed as writers in their own right, and how their self-efficacy impacted their ability to assist their young elementary students to foster and support creativity. Using data collected from questionnaires, personal histories, and interviews, Cremin classified the 16 teachers according to their writing profiles. She then selected three participants: one who expressed a highly positive level of self-efficacy, one who expressed a low level of self-efficacy, and one with a mid-range level of self-efficacy. When the three teachers were asked to produce an original narrative, each reported anxieties and uncertainties. None had completed a narrative since their own school days. According to Cremin, all participants persevered, took risks, and completed the task that was later published. At the end of the two-year study, teachers expressed satisfaction in the process and the finished product, even though at the time it involved considerable discomfort. The three teachers Cremin reported on in detail, along with the other 13 who participated in the story writing task, all reported that they risked modeling writing to their students and/or regularly composed alongside their students. Cremin suggested that teachers who engaged in composition at their own level underwent a conversion of sorts that enabled them to better direct and relate to their students as they engaged in writing tasks.

According to Faigley, Daly, & Witte (2001) individuals with high levels of writing apprehension tended to avoid writing tasks. Faigley, Daly, and Witte reported that people who

experienced high apprehension levels found writing difficult, unrewarding, and at some levels frightening. Proponents of holistic scoring reported on the serious reliability problems it engenders (Cooper & Odell, 1977; White, 1993). "To overcome it, groups of teachers or researchers have to work together to train themselves as raters" (Cooper & Odell, 1977, p.21). When teachers develop their understanding of the writing task, they are better able to guide their students through the writing process (Fisher, Brooks, and Lewis, 2002).

Firestone et. al. (2002) used surveys, observations and interviews to examine how test preparation affected math and science instructional practices in New Jersey schools and found they had "both good news and bad news" to report. Positives or "good news" included: test preparations that engaged students in a variety of tasks; the use of new instructional approaches that included problem solving; the use of higher order questioning techniques; and a movement away from competition. Negatives or "bad news" included: an appearance of no clear focus among topics; superficial attempts to deal with a new type of test; and the persistence of drill-oriented approaches in relatively "poor" urban areas.

> Taken together, our findings suggest that state testing is neither the magic policy bullet that advocates of accountability hope for nor the force for deskilling, dumbing down, and disparity of life chances that certain opponents have claimed. On the basis of what we have observed, the best that can be said for state testing is that, when properly designed, it can sensitize teachers to new instructional approaches and promote shifts in the content that is taught. (Firestone et. al., 2002, p. 1518)

Personal Involvement

Prompted by the concerns and criticisms of colleagues who attacked holistic scoring practices "on the grounds that it prevented raters from fully interacting with student writing" (Huot, 1993, p. 227), Huot examined the procedures associated with holistic scoring. He hypothesized that the fast-paced reading process hindered raters' personal involvement and thereby impeded judgments about the quality of the essay. Using two sets of practicing English teachers as essay readers, one trained and one untrained, and 24 essays that had been previously scored during a task force project, Huot compared the two groups' responses during and after scoring sessions. Huot found that both novice and expert groups responded personally to the students' writings. He also noted that the use of the holistic scoring rubric made it easier for both novice and expert groups to score the essays and agree with each other.

Pula and Huot (1993) replicated the study conducted by Huot (1993) and extended it with additional data collected during post-scoring interviews. Pula and Huot (1993) were interested in studying the influences of background and personal experience on the scoring process. In the focused interview sessions, Pula and Huot encouraged raters to "tell their own stories." Of the eight raters that were interviewed, five recalled deficits in their writing performances that were later overcome, sometimes through instruction, sometimes though help from knowledgeable individuals. The raters reported that they recognized similar patterns of difficulties in the essays they scored. "Some indicated that they downgraded essays for insufficiencies in these areas, (Pula & Huot, 1993, p. 249)" while others were more lenient, understanding that what they had overcome would be similarly overcome by the student writers and that the student writers, like them, were "just in a temporary stage" (p. 249).

Raters and scorers bring a unique background and outlook to the evaluation process (Maki, 2004; Pula & Huot, 1993; White, 1993). Training the raters, using multiple raters, and using multiple assessments are all valid tactics that can be used to enhance reliability, ensure fairness, and control for human factors. It has been well argued that training can improve the reliability of essay scores (Maki, 2004), and while personal involvement is necessary to the reading process (Hout, 1993), evaluators must strive to make sound judgments that are unhampered by personal background and experience issues. White (1986) stated:

> Teachers who are excessively rigid or insecure often have difficulty adopting group standards, and faculty members who take pride in their differences with their colleagues may resent the entire process. But even such apparently unsuited teachers turn out to be delightful at readings, while some well-recommended people read erratically and inattentively. (p. 71)

School administrators and teachers can support student growth and improvement in writing by maintaining clear goals for assessing student work (National Writing Project & Nagin, 2003). Whatever assessment instruments are used for writing, there must be explicit connections among curricular aims, standards, instructional needs, the tests, and their scoring criteria or rubrics. Most importantly, assessment should have an instructional purpose, not just and evaluative or administrative one (Higgins, Miller, & Wegmann, 2007; National Writing Project & Nagin, 2003).

CHAPTER THREE: METHODOLOGY

Inter-rater reliability is the extent to which different raters assign the same score to the same writing sample (White, 1993). This study examined the inter-rater reliability of scores assigned by 17 raters prior to, during, and after three school-based training sessions aimed at improving inter-rater reliability. This study tested the hypothesis that scoring consistency, or inter-rater reliability, improves with training and discussion. Additionally, a qualitative study of the relationship between teachers' writing confidence levels [as measured by the *Teacher/Writer Questionnaire* (Bowie, 1996)], the assignment of scores, and response to training was also conducted.

Participants

Training Participants

The teachers who received the training were all faculty members at one elementary school which will be referred to as Claver Elementary. The school name and all the participants' names are pseudonyms. The school was located in a middle-class suburban area in Florida. Seven participants were assigned to third grade, six were assigned to fourth grade, and four were assigned to fifth grade. Ten participants had bachelor's degrees, and seven had masters. Years of experience ranged from 1 to 20 years, with 5 teachers having less than 5 years of experience. Twenty teachers participated in at least two out of the three training sessions, and seventeen teachers fully participated in all three sessions.

Training Presenters

The presenters were faculty members of the Claver Elementary School. They conducted the training sessions that included a review of the FCAT writing rubric and follow-up scoring/discussion sessions. One of the presenters, Barbara, had 14 years of teaching experience. She had conducted five writing workshops/training sessions in the county where she worked prior to her transfer to Claver Elementary. The second presenter, Jean, was a National Board certified teacher with a Middle Childhood Generalist specialization. She had 14 years of teaching experience and had conducted informal action research studies of student responses to on-demand writing prompts. Both presenters displayed an interest in conducting a training activity designed to help third, fourth, and fifth-grade teachers better understand on-demand writing and assessment.

Writing Samples

The writing samples used in this study were responses made by fourth-grade students to expository and narrative prompts during pilot and calibration studies conducted by the Florida Department of Education's (Florida Department of Education computer software, 2006 & 2007). A total of four prompts, two expository and two narratives, were used. The prompts can be found in Table 1.

Table 1 **Writing Prompts**

Writing to Explain (Expository)	Writing to Tell a Story (Narrative)
Students were directed to think about something that is special to them and explain why it is special.	Students were directed to write a story about a field trip to a special place.
Students were asked to choose something fun to do outside and explain what makes that activity fun.	Students were asked to write a story about a time an animal does something smart.

Procedures

The study focused on teachers' evaluations of student responses to on-demand expository and narrative writing prompts. All writing samples used in this study were secured from *FCAT Writing+ Training Materials: Anchor Papers and Qualification Sets* (computer software, 2006 & 2007) provided to all Florida schools by the Florida Department of Education. A total of twenty-two writing samples were used during the pre-test, post-test, and practice scoring/discussion sessions. Before the study was conducted, the researcher submitted the research protocol to the University of Central Florida Institutional Review Board (UCF IRB) for review and approval. The notice of approval can be found in Appendix G.

Training Prep

Prior to the training sessions, the presenters met with the researcher to create an agenda for the three training sessions. After considering the needs and varied experiences of the staff, it was decided that a basic review of the four areas assessed on the FCAT writing test (focus organization, support, and conventions) was appropriate. Overhead transparencies were

developed to guide the presentation and to provide visual cues. Jean agreed to provide excerpts from 4th grade student writings to use as examples. The introduction and general review of the scoring criteria took place during Session 1.

The presenters and researcher also outlined procedures to be followed during Sessions 2 and 3. The main foci of Sessions 2 and 3 were scoring practice and small-group discussions. During Sessions 2 and 3, Barbara and Jean's roles were that of observers and moderators, they did not actively participate in or contribute to the group decisions.

Pre-Training Survey

Prior to the training sessions, teachers completed the *Teacher/Writer Questionnaire* (Bowie, 1996) designed to gather information about their attitudes and feelings related to writing tasks and assessments. The surveys were distributed through the school mail system two weeks prior to training, and all were completed and returned prior to Session 1. The questionnaire consisted of 40 questions that were answered using a five-point Likert scale. Five demographic questions aimed at identifying respondents' teaching assignments, education, and years of experience were added for purposes of the study. The survey statements and responses can be found in Appendix A.

Pre-Test

Prior to participating in the in-house training program, teachers were asked to score four writing samples, two expository and two narratives. The four papers assessed during the pre-test were selected from anchor papers and qualification sets distributed by the Florida Department of Education (Florida Department of Education computer software, 2006 & 2007). Teachers were

instructed to score the writing samples using the FCAT Writing+ rubric (Appendix C). Each teacher received a copy of the rubric to use during the evaluation. Teachers met at one location to score the writing samples, and did not consult with each other during the preliminary assessment process.

The holistic FCAT Writing+ rubric used in the evaluation had a score range from 0-6, with a 6 being the highest. The writing samples had been previously scored by the Florida Department of Education's testing service and were representative of Levels 3, 4, and 5 of the FCAT Writing+ scale. Levels 1, 2, and 6 were not used for the pre-training and post-training assessments because Levels 1 and 2 were the minimal levels and Level 6 was the highest. Since papers at those three levels were at the extreme ends of the spectrum, they were not likely to generate scores both above and below the scores assigned by the Florida Department of Education. For example, a Level 6 paper could only be rated at or below Level 6. Conversely, Levels 3, 4, and 5 were used because they were considered mid-level papers, and the teachers at Claver reported that those levels were more difficulty to grade.

Session 1

The two presenters introduced the Florida Department of Education's writing rubric and explained the distinctions between the six achievement levels. While reviewing the various levels of the rubric, the presenters attempted to identify and clarify vague terms and/or "gray" areas. For example, they explained the differences between extensions and elaborations, which both fall into the "support" category. They also defined bare ideas and gave examples of writings that were "off topic." After the rubric presentation, teachers met in small groups of four or five,

to identify concepts that they continued to find confusing or unclear. One member from each group recorded the concerns of the group and delivered them in writing to the presenters.

Session 2

The two presenters began Session 2 with a 5 minute presentation aimed at clarifying issues and concerns noted during the Session 1 discussions. After the short recap and discussion, teachers participated in scoring activities both individually and as members of small groups. Group assignments were randomly determined prior to Session 2. Name cards were used to direct the participants to their assigned groups. Procedures for Session 2 are outlined below:

1. Each teacher received a copy of the holistic writing rubric, a form to record scores and rationale, and seven writing samples labeled with the letters A-G. The papers selected for the scoring/discussion session were selected from anchor papers and qualification sets distributed by the Florida Department of Education (Florida Department of Education computer software, 2006 & 2007). The writing samples represented all six levels of the writing rubric and were arranged on a clipboard in alphabetical order.
2. Each group was given three colored flags, red, yellow, and green, along with a flag holder. The flags were approximately eight inches tall and made of a stiff felt material. The flags represented three levels of group activity.
3. The red flag was displayed to indicate that the group was engaged in the initial period of silent reading/scoring of the first prompt.
4. When all members indicated that initial scoring was complete, the green flag was displayed and the group engaged in open discussion as they attempted to arrive at a "consensus" score.

5. When the group reached a consensus, the yellow flag was displayed, and when all groups displayed yellow flags, the presenters asked that each group announce their "consensus" score.
6. After all scores were announced, the presenters announced the score the Florida Department of Education (FLDOE) assigned the particular writing sample. The presenters followed the announcement with a reading of the rationale for the FLDOE score.
7. The process of individual scoring, group discussion to form consensus, the announcements of the groups' scores, and the announcement of the FLDOE score and rationale was repeated for each of the remaining six writing samples.

Session 3

With the exception of the opening presentation, the procedure outlined in Session 2 was repeated. Session 3 began with the presenters reviewing the procedures for individual and group scoring. Group assignments were randomly determined prior to Session 3, and name cards were again used to direct the participants to their assigned groups. Writing samples for Session 3 were labeled H-N. As in Session 2, teachers recorded their initial and group scores and the rationale for those scores on individual forms/charts.

Post-Test

At the end of Session 3, teachers were asked to score four writing samples, two expository and two narratives. The four papers assessed during the post-test were selected from anchor papers and qualification sets distributed by the Florida Department of Education (Florida

Department of Education computer software, 2006 & 2007). The four papers selected for the post-test were determined by the Florida Department of Education to be representative of Levels 3, 4, and 5. Teachers scored the writing samples using the FCAT Writing+ rubric. Each teacher received a copy of the rubric to use during the evaluation. Teachers did not consult with each other during the final assessment process.

Time Elements

Each of the three training sessions lasted 60 – 70 minutes. There were two-week breaks between sessions. The pre-test took place at the beginning of Session 1. The participants were given no time constraints, and that process took no more than 15 minutes. The post-test took place at the end of Session 3. Again, no time restraints were involved, and the process did not exceed 15 minutes.

Assumptions and Limitations

- The third, fourth, and fifth-grade teachers participating in this study had a basic knowledge of writing competencies measured by the holistic rubric.
- The study was limited to the teachers at one elementary school; therefore, generalizations to other schools of different demographics would not be appropriate.
- The study was limited to one holistic scoring rubric designed to evaluate how well writers integrate four elements: focus, organization, support, and conventions; therefore, generalizations to other rubrics and assessment methods would not be appropriate.

- The study was limited to a small number of participants and a small number of scoring samples; therefore, all conclusions relate only to the school in the study and are necessarily tentative.
- The responses to the questionnaire can only represent the presumed honesty of the respondents, as this represents self-reported data.

CHAPTER FOUR: RESULTS

This study examined the inter-rater reliability of writing scores assigned by 17 teachers as part of a school-based training program. The analyses included *t*-tests that compared the participants' mean scores to the scores assigned by the Florida Department of Education (FLDOE), a within-group analysis of reliability as measured by Cronbach's Alpha, and percentage agreement analyses. This study tested the hypothesis that scoring consistency, or inter-rater reliability, improves with training and discussion.

A qualitative component focused on the scoring patterns of six participants. The purpose was to study the relationship between teachers' writing confidence levels [as measured by the *Teacher/Writer Questionnaire* (Bowie, 1996)], the assignment of scores, and response to training. Twelve items on the Teacher/Writer Questionnaire were statements that solicited information about respondents' feelings of self-efficacy related to writing and writing tasks. Responses to those items were used to assign a "confidence" score and rank to the participants (Appendix B). A study of the scoring patterns of the three targeted participants identified as having the highest confidence levels and the three with the lowest confidence levels was conducted in an attempt to determine whether their perceptions of their personal writing ability influenced their assessment of students' writing.

Pre-Test Score Analyses

T- Tests

One-sample *t*-tests were conducted on the writing evaluations completed by the 17 participants prior to and following the training sessions. The purpose was to compare the mean scores of the study participants to the scores assigned by the testing department of the Florida

Department of Education (FLDOE). Accordingly, the null hypothesis in each case was that there was not a statistically significant difference between the mean score of the participant group and the FLDOE mean. Writing samples A, B, C, and D were assessed prior to the training sessions. A summary of the means, standard deviations, and standard error of the means can be found in Table 2.

Table 2 **Pre-Test Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
Pre-test expository A	17	3.53	.874	.212
Pre-test expository B	17	2.53	1.068	.259
Pre-test narrative C	17	3.94	1.088	.264
Pre-test narrative D	17	3.12	.600	.146

There were four writing samples, two expository and two narratives used in the pre-test. According to the FLDOE raters, Expository Response A was Level 4, Expository Response B was Level 3, Narrative Response C was Level 5, and Narrative Response D was Level 4. Tables 3-6 summarize results of *t*-tests conducted to compare the mean scores of the participants to the scores assigned by the Florida Department of Education.

Table 3 **One-Sample *t*-Test: Writing Sample A (Pre-Test)**

	Test Value = 4					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Pre-test Expository A	-2.219	16	.041	-.471	-.92	-.02

Table 3 summarized the results of the *t*-test conducted to determine whether the study participants' mean score for Expository Prompt Response A differed from the score of 4 assigned by the Florida Department of Education. The test was conducted using an alpha of .05. The participants' mean score (M = 3.53, s = .874, SEM = .212) was statistically significantly different from 4, t (16) = -2.22, $p < .05$. The 95% confidence interval for the mean difference was -.92 to -.02, and it did not contain zero. In that case, the null hypothesis was rejected. The results provided evidence that there was a statistically significant difference in the participants' mean score for Expository Prompt Response A and the FLDOE score.

Table 4 **One-Sample *t*-Test: Writing Sample B (Pre-Test)**

	Test Value = 3					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Pre-test expository B	-1.817	16	.088	-.471	-1.02	.08

Table 4 summarized the results of the *t*-test conducted to determine whether the study participants' mean score for Expository Prompt Response B differed from the score of 3 assigned

by the Florida Department of Education. The test was conducted using an alpha of .05. The participants' mean score (M = 2.53, s = 1.068, SEM = .259) was not statistically significantly different from 3, *t* (16) = -1.82, p > .05. The 95% confidence interval for the mean difference was -1.02 to .08, and it did contain zero. In this case, the results provided evidence that there was not a statistically significant difference in the participants' mean score for Expository Prompt Response B and the FLDOE score.

Table 5 **One-Sample *t*-Test: Writing Sample C (Pre-Test)**

	Test Value = 5					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Pre-test narrative C	-4.012	16	.001	-1.059	-1.62	-.50

Table 5 summarized the results of the *t*-test conducted to determine whether the study participants' mean score for Narrative Prompt Response C differed from the score of 5 assigned by the Florida Department of Education. The test was conducted using an alpha of .05. The participants' mean score (M = 3.94, s = 1.088, SEM = .264) was statistically significantly different from 5, *t* (16) = -4.01, p < .05. The 95% confidence interval for the mean difference was –1.62 to - .50, and it did not contain zero. In that case, the null hypothesis was rejected. The results provided evidence that there was a statistically significant difference in the participants' mean score for Narrative Prompt Response C and the FLDOE score.

Table 6 **One-Sample *t*-Test: Writing Sample D (Pre-Test)**

	Test Value = 4					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Pre-test narrative D	-6.061	16	.000	-.882	-1.19	-.57

Table 6 summarized the results of the *t*-test conducted to determine whether the study participants' mean score for Narrative Prompt Response D differed from the score of 4 assigned by the Florida Department of Education. The test was conducted using an alpha of .05. The participants' mean score (M = 3.12, s = .600, SEM = .146) was statistically significantly different from 4, t (16) = -6.06, $p < .05$. The 95% confidence interval for the mean difference was –1.19 to - .57, and it did not contain zero. In that case, the null hypothesis was rejected. The results provided evidence that there was a statistically significant difference in the participants' mean score for Narrative Prompt Response D and the FLDOE score.

In sum, there was a statistically significant difference between the participants' mean scores and the FLDOE scores in 3 out of 4 cases. There was significant agreement between the score of 3 assigned by the Florida Department of Education and the participants' mean score for Expository Prompt Response B.

Intraclass Correlations

Cronbach's Alpha was the statistic used to assess inter-rater reliability in accordance with guidelines recommended by Atkinson and Murray (1987). A two-way mixed-model intraclass

correlation was used because this study compared multiple raters that scored the same writing samples. For the four pre-test scores, the alpha coefficient was .913, well above the .80 standard for high intraclass correlations (Cherry & Meyer, 1993). However, the high alpha coefficient should be interpreted with caution because researchers have noted that reliability coefficients increase as the number of raters increase (Moore & Young, 1997; Cherry & Meyer, 1993).

Pre-Test Percentage Agreement Analysis

An evaluation of the pre-test percentage agreement was conducted and is outlined in Table 7. Table 7 shows percentages for the number of times the participants' scores were in complete agreement with or within one point of the FLDOE score for each of the four pre-test samples

Table 7 Pre-Test Percentage Agreement - Participants' Scores and FLDOE

	Number of scores	Number in exact agreement	Percentage in exact agreement	Number within 1 score point	Percentage within 1 score point
Expository Sample A	17	7	41.17	15	88.24
Expository Sample B	17	4	23.53	14	82.35
Narrative Sample C	17	5	29.41	10	58.82
Narrative Sample D	17	4	23.53	15	88.24
Total	68	20	29.41	54	79.41

On the pre-test, scores assigned by the study participants were in exact agreement with the FLDOE in 20 out of 68 cases, or 29.41% of the time. They were within one point of the FLDOE in 54 out of 68 cases, or 79.41% of the time.

Survey Responses

The seventeen teachers who participated in the study also responded to a 40-item Teacher/Writer Questionnaire (Bowie, 1996). The survey results can be found in Appendix A. It lists the frequencies of responses in each category as well as the mean scores. The 40-item Teacher/Writer Questionnaire used a 5-point Likert scale. Bowie (1996) conducted a test of the questionnaire's internal consistency using a Cronbach Alpha. That test produced a .94 reliability coefficient for the Teacher/Writer Questionnaire.

Survey respondents indicated strong agreement with the statements by circling the number 1, agreement by circling the number 2, uncertainty by circling the number 3, disagreement by circling the number 4, and strong disagreement by circling the number 5. Of the 40 statements, 23 were worded positively, and 17 were worded negatively. A low mean for a positively worded statement indicates a positive attitude towards that statement. For negatively worded statements a low mean indicates a more negative attitude.

The 40-question Teacher/Writer Questionnaire used in this study was piloted and developed by Bowie (1996). Twelve items on the Teacher/Writer Questionnaire explored the respondents' feelings of self-efficacy related to writing and writing tasks. Responses to those items (2, 7, 11, 15, 20, 22, 30, 32, 34, 35, 37, and 39) were used to assign a "confidence" score and rank to the respondents. Since items 22, 30, 34, 35, 37, and 39 were negatively worded the responses were reversed or recoded to determine individual scores. Agreement with items 2, 7,

11, 15, 20, and 32 and disagreement with items 22, 30, 34, 35, 37, and 39 would indicate strong levels of self-confidence or self-efficacy. Conversely, disagreement with items 2, 7, 11, 15, 20, and 32 and agreement with items 22, 30, 34, 35, 37, and 39 would indicate lower levels of self-confidence or self-efficacy. After recoding, a test of internal reliability was conducted. The reliability alpha coefficient for the twelve "confidence" items was .93, a very high rating.

Confidence Rankings

The study participants were ranked according to their responses, from those with higher confidence levels to those with lower confidence levels. Responses to the survey were collected anonymously; therefore, participants' names were assigned for the sake of discussion and cannot be interpreted as an indication of gender. Participants Alice, Betty, and Calvin had the lowest scores and therefore the highest reported confidence levels. Participants Opal, Patrick, and Quincy had the highest scores and therefore the lowest reported confidence levels. A summary of the ranks of all participants can be found in Appendix B.

"High-Confidence" and "Low-Confidence" Groups Compared

Means

The pre-test and post-test scores assigned by the three teachers with the highest confidence scores and the three teachers with the lowest confidence scores are listed in Table 8. Also shown in Table 8 are the scores assigned by the Florida Department of Education (FLDOE).

Table 8 "High" and "Low" Confidence Groups Pre-Test and Post-Test Scores

	Group	N	Mean	Std. Deviation	Std. Error Mean	FLDOE Score
Pre-test expository A	High	3	3.00	1.000	.577	4
	Low	3	4.33	.577	.333	4
Pre-test expository B	High	3	2.33	1.155	.667	3
	Low	3	3.33	1.155	.667	3
Pre-test narrative C	High	3	3.67	1.155	.667	5
	Low	3	4.33	1.155	.667	5
Pre-test narrative D	High	3	3.33	.577	.333	4
	Low	3	3.33	.577	.333	4
Post-test expository E	High	3	4.67	1.528	.882	5
	Low	3	5.67	.577	.333	5
Post-test expository F	High	3	3.33	.577	.333	4
	Low	3	3.67	.577	.333	4
Post-test narrative G	High	3	4.33	.577	.333	4
	Low	3	3.33	.577	.333	4
Post-test narrative H	High	3	3.67	1.528	.882	3
	Low	3	3.33	.577	.333	3

On all pre-training assessments, the mean score of the high confidence group is less than the score assigned by the FLDOE. The mean scores of the high confidence group is less than the mean score of the low confidence group in 3 out of 4 cases and is equal to the mean score of the low confidence group in 1 out of 4 cases. On the pre-test the mean score of the low confidence

group is higher than the FLDOE score in 2 out of 4 cases and lower than the FLDOE score in 2 out of 4 cases.

On the post-training assessments, the mean scores of both the high confidence group and the low confidence group is higher than the FLDOE score in 2 out of 4 cases and lower than the FLDOE score in 2 out of 4 cases. The mean scores of the high confidence group is less than the mean score of the low confidence group in 2 out of 4 cases and higher than the mean score of the low confidence group in 2 out of 4 cases.

Percentage Agreement

Florida DOE writing evaluations that are within 1 point of each other are reported as the average of the two scores (*Florida Department of Education FCAT Handbook*, 2005). Therefore, in this study, consideration is given to how often that condition applies to the scores of the "high" and "low" confidence groups. This was done by looking at instances where the standard deviations were greater than 1. On the pre-training evaluations the standard deviation was greater than or equal to 1 in five instances: 3 instances from the "high-confidence" group and 2 instances from the "low-confidence" group. On the post-training evaluations the standard deviation was greater than or equal to 1 in only two instances: both from the "high-confidence" group. An evaluation of the agreement percentages for "high" and "low" confidence groups and the FLDOE follows.

Table 9 "High" and "Low" Confidence Groups Compared to FLDOE Percentage Agreement for Pre-Test and Post-Test

	Group	Number of scores	Number in exact agreement	Percentage in exact agreement	Number within 1 score point	Percentage within 1 score point
Pre-test Samples	High	12	5	41.67	8	66.67
Post-test Samples	High	12	4	33.33	10	83.33
Pre-test Samples	Low	12	5	41.67	11	91.67
Post-test Samples	Low	12	6	50.00	12	100.00

On the pre-test, scores assigned by the "high-confidence" participants were in exact agreement with the FLDOE in 5 out of 12 cases, or 41.67% of the time. They were within one point of the FLDOE in 8 out of 12 cases, or 66.67% of the time. On the post-test, scores assigned by the "high confidence" participants were in exact agreement with the FLDOE in 4 out of 12 cases, or 33.33% of the time. They were within one point of the FLDOE in 10 out of 12 cases, or 83.33% of the time. While the percentage of exact agreement between the scores assigned by the "high-confidence" group and the FLDOE scores declined on the post-test, a higher percentage of the those scores were within one score point of each other.

On the pre-test, scores assigned by the "low confidence" participants were in exact agreement with the FLDOE in 5 out of 12 cases, or 41.67% of the time. They were within one point of the FLDOE in 11 out of 12 cases, or 91.67% of the time. On the post-test, scores assigned by the "low confidence" participants were in exact agreement with the FLDOE in 6 out

of 12 cases, or 50% of the time. They were within one point of the FLDOE in 12 out of 12 cases, or 100% of the time. For the "low-confidence" group, both the percentage of exact agreement and the percentage of scores within one score point increased on the post-test.

Agreement Within Groups

A further study of the evaluation patterns of the high and low confidence groups was conducted using scores assigned to writing samples during the group sessions. During those sessions a total of 14 samples were evaluated, 7 during Training Session 2 and 7 during Training Session 3. The 14 scores assigned by the three "high confidence" individuals and 14 scores assigned by the three "low confidence" individuals during Training Sessions 2 and 3 were separately assessed for inter-rater reliability using a mixed model intraclass correlation. The alpha coefficient for the "high confidence" group was .886. The alpha coefficient for the "low confidence" group was slightly higher at .936.

A summary of the "high-confidence" and "low confidence" groups' mean scores for Session 2 is shown in Table 10. In that table the writing samples are listed in the order they were evaluated.

Table 10 "High" and "Low" Confidence Groups' Statistics: Scores Assigned in Session 2

Group	Sample	N	Mean	Std. Deviation	Std. Error Mean
High	A1	3	2.67	1.528	.882
Low		3	2.33	.577	.333
High	B2	3	2.33	.577	.333
Low		3	2.00	1.000	.577
High	C3	3	5.33	.577	.333
Low		3	5.00	.000	.000
High	D4	3	1.67	.577	.333
Low		3	2.33	.577	.333
High	E5	3	2.67	.577	.333
Low		3	2.67	.577	.333
High	F6	3	3.33	.577	.333
Low		3	4.00	1.000	.577
High	G7	3	4.33	.577	.333
Low		3	3.33	.577	.333

During Training Session 2, the first group session, highly confident participants' mean scores differed by more than one point in only 1 of the 7 cases. The less confident participants were in full agreement on writing sample C, and their scores differed by more than one point in 2 of the 7 cases.

Table 11 summarizes the "high-confidence" and "low confidence" groups' mean scores for Session 3. Writing samples are listed in the order they were evaluated.

Table 11 "High" and "Low" Confidence Groups' Statistics: Scores Assigned During Session 3

Group	Sample	N	Mean	Std. Deviation	Std. Error Mean
High	H8	3	3.67	1.155	.667
Low		3	4.33	.577	.333
High	I9	3	2.67	.577	.333
Low		3	2.33	.577	.333
High	J10	3	2.67	.577	.333
Low		3	2.33	.577	.333
High	K11	3	2.00	1.000	.577
Low		3	1.33	.577	.333
High	L12	3	6.00	.000	.000
Low		3	6.00	.000	.000
High	M13	3	2.00	1.000	.577
Low		3	2.67	.577	.333
High	N14	3	3.33	1.155	.667
Low		3	4.00	.000	.000

During Training Session 3, the second group session, the 6 focus teachers' scores were the same for Writing Sample L. The highly confident participants' mean scores differed by more than one point in 4 of the other 6 cases. The less confident participants were in full agreement on Writing Sample N, and their mean scores differed by no more than one point in the remaining 5 cases.

Individual Scores Compared to Training Group Consensus Scores

A comparison of individual scores to group consensus scores revealed the following: In Training Session 2, the number of times the individual scores of the highly confident participants differed from their group consensus scores was equal to the number of times scores of less confident participants differed from their group consensus scores: 6 out of 21 cases or 31% of the cases. The directional value of the change was also equally split. Both groups changed their scores from a higher score to a lower score in 3 cases and from a lower score to a higher score in 3 cases. See Table 12.

Table 12 "High" and "Low" Confidence Groups' Score Changes (Session 2)

ID	Group	Group consensus score minus individual score						
Alice	Hi	-1	1	0	0	-1	0	0
Betty	Hi	1	0	0	0	0	0	-1
Calvin	Hi	0	0	0	0	0	1	0
Opal	Low	0	-1	0	0	0	-1	0
Patrick	Low	0	0	0	0	1	0	1
Quincy	Low	0	0	0	-1	0	0	1

In Training Session 3, the number of times the individual scores of the highly confident participants differed from their group consensus scores was 7 out or 21 or 33% of the cases. Three of the individual scores were higher than the group consensus, and four were lower than the group consensus. The individual scores of the less confident group differed from their group

consensus scores in 4 out of 21 cases or 19% of the cases. Two of the individual scores were higher than the group consensus, and two were lower than the group consensus. See Table 13.

Table 13 "High" and "Low" Confidence Groups' Score Changes (Session 3)

ID	Group	Group consensus score minus individual score						
Alice	Hi	1	-1	-1	-1	0	0	0
Betty	Hi	1	0	0	0	0	1	1
Calvin	Hi	0	0	0	0	0	0	0
Opal	Low	1	0	0	0	0	-1	0
Patrick	Low	-1	0	0	1	0	0	0
Quincy	Low	0	0	0	0	0	0	0

In sum, there was little difference between high and low confidence groups in Session 2 in regards to how frequently their individual scores differed from the group consensus scores. It was also noted that the group consensus score was a 1 point increase over the individual score in an equal number of cases as it was a 1 point decrease from the individual score. In Session 3, there was a slight increase in the number of times the individual scores of members of the high confidence group differed from the consensus scores and a slight decrease in the number of times the individual scores of the low confidence group differed from the consensus score. For both groups, it was noted that the group consensus score was a 1 point increase over the individual score in a nearly equal number of cases as it was a 1 point decrease from the individual score.

Qualitative Analyses

In the scoring sessions conducted during Training Sessions 2 and 3, participants were asked to provide rationales for the scores they assigned to 14 writing samples. Those rationales were written prior to group discussions. A qualitative analysis of the rationales provided by the "high confidence" group and "low confidence" group follows.

Participants' Education and Experience

A summary of the education and experience of the individuals with "high" and "low" confidence levels is shown in Table 14. There were no significant patterns noted in the demographics related to education and experience. Individuals in both groups had varying amounts of experience and educational achievement levels. However, all the highly confident teachers were assigned to grade 4.

Table 14 Education and Experience - "High" and "Low" Confidence Groups

Group	Name	Current Grade	Teaching Experience	Highest Degree
High	Alice	4	17 years	Bachelors
High	Betty	4	17 years	Masters
High	Calvin	4	2 years	Bachelors
Low	Opal	5	1 year	Bachelors
Low	Patrick	3	12 years	Masters
Low	Quincy	5	20 years	Masters

Method for Reporting Rationale

During Sessions 2 and 3, the presenters requested that participants note the rationale for their scores on a chart they provided (Appendix D). They did this so the participants would have a point of reference during group discussion. During this process, the participants were allowed to use the following codes to expedite the assessment process.

- F = Focus
- O = Organization
- BI = Bare Ideas
- EX = Extensions
- EL = Elaborations
- V = Voice
- ML = Mature Language
- C = Conventions

Examples of scores and the related rationales of individuals, groups, and the FLDOE can be found in Appendixes E and F. Summaries of the number of positive and negative comments made as part of the scoring rationale are shown in Tables 15 and 16. Also noted in Tables 15 and 16 are the number of times the participants failed to note any rationale for the scores assigned.

Table 15 **Comment Analysis: "High-Confidence" Group**

	Number of positive comments	Number of negative comment	Number of times no rationale noted
Alice	6	9	5
Betty	17	16	0
Calvin	16	1	6
Group Totals	39	26	11

Table 15 showed that the "high-confidence" group made a total of 39 positive comments and a total of 26 negative comments. Additionally, participant Calvin accounted for 41% of the positive comments and only 4% of the negative comments. Among the "highly-confident" participants, there were a large number of instances where no rationale was given for the assigned score. This occurred in 11 out of 42 cases, or 26% of the total number of opportunities the "high-confidence" group had to note rationale. This is a relatively high percentage when compared to the whole group. The seventeen participants had a total of 238 opportunities to note a rationale for their assigned scores. There were 33 instances in the entire population of this study where no rationale was noted, or 14% of the total. The "high-confidence" group accounted for 11 of those cases, which was 5% of the total.

Table 16 **Comment Analysis: "Low-Confidence" Group**

Participant	Number of positive comments	Number of negative comment	Number of times no rationale noted
Opal	19	6	0
Patrick	14	12	0
Quincy	41	5	4
Group Totals	74	23	4

Table 16 showed that the "low-confidence" group made a total of 74 positive comments and a total of 23 negative comments. Additionally, participant Quincy accounted for 55% of the positive comments and 22% of the negative comments. The "less-confident" participant, failed to note a rationale for their assigned scores in 4 out of 42 cases, or only 10% of the total number of opportunities the "low-confidence" group had to note rationale. When compared to the whole group, this accounted for only 2% of the total.

Two Outliers

As was indicated in the "high-confidence" and "low-confidence" ranking process, (Appendix B) there was a considerable difference between the confidence scores of the top ranked participant and the lowest ranked participant. The individual with the highest confidence rank had a score of 17, and there was a 6-point difference between that individual and the second highest ranked individual who had a score of 23. Similarly, the lowest ranked participant had a score of 51, and there was a 7-point difference between that score and the next lowest score, 44.

Since these two individual scores stand out from the pack, a closer evaluation of those individuals' score patterns was appropriate.

It was noted that in Training Session 2, participant Alice, the individual with the highest confidence level, accounted for 3 of the 6 scores (50%) that differed from the group consensus (Table 12). In Training Session 3, Alice's scores accounted for 4 of the 7 scores (57%) that differed from the group consensus (Table 13).

Alice also made the smallest number of positive comments, a total of six. She was the only one of the six participants whose negative comments outnumbered the positive comments. She made nine negative comments and failed to give any rationale for the assigned score in 5 out of 14 cases (Table 15).

In contrast, participant Quincy, the individual with the lowest confidence level, accounted for 2 of the 6 scores (33%) that differed from the group consensus in Training Session 2. In Training Session 3, Quincy's scores accounted for 0 of the 4 (0%) that differed from the group consensus (Table 13).

Quincy made the largest number of positive comments, a total of 41. His positive comments outnumbered all of the other positive responses totals by more than 2 to 1. Quincy made five negative comments and failed to give any rationale for the assigned score in 4 out of 14 cases (Table 15).

Post-Test Score Analyses

T-Tests

Writing samples E, F, G, and H were assessed after Training Session 3. A summary of the means, standard deviations, and standard error of the means can be found in Table 17.

Table 17 **Post-Test Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
Post-test expository E	17	4.82	.951	.231
Post-test expository F	17	3.41	.618	.150
Post-test narrative G	17	3.71	.686	.166
Post-test narrative H	17	3.53	.800	.194

There were four writing samples, two expository and two narratives used in the post-test. According to the FLDOE raters, Expository Response E was Level 5, Expository Response F was Level 4, Narrative Response G was Level 4, and Narrative response H was Level 3. Tables 18-21 summarize results of *t*-tests conducted to compare the mean scores of the participants to the scores assigned by the Florida Department of Education.

Table 18 **One-Sample *t*-Test: Writing Sample E (Post-Test)**

	Test Value = 5					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Post-test expository E	-.765	16	.455	-.176	-.67	.31

Table 18 summarized the results of the *t*-test conducted to determine whether the study participants' mean score for Expository Prompt Response E differed from the score of 5 assigned by the Florida Department of Education. The test was conducted using an alpha of .05. The participants' mean score (M = 4.82, s = .951, SEM = .231) was not statistically significantly different from 5, *t* (16) = -.765, p > .05. The 95% confidence interval for the mean difference was -.67 to .31 and it did contain zero. In this case, the results provided evidence that there was not a statistically significant difference in the participants' mean score for Expository Prompt Response E and the FLDOE score.

Table 19 One-Sample *t*-Test: Writing Sample F (Post-Test)

	Test Value = 4					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Post-test expository F	-3.922	16	.001	-.588	-.91	-.27

Table 19 summarized the results of the *t*-test conducted to determine whether the study participants' mean score for Narrative Prompt Response F differed from the score of 4 assigned by the Florida Department of Education. The test was conducted using an alpha of .05. The participants' mean score (M = 3.41, s = .618, SEM = .150) was statistically significantly different from 4, *t* (16) = -3.922, p < .05. The 95% confidence interval for the mean difference was –.91 to - .27, and it did not contain zero. In that case, the null hypothesis was rejected. The

results provided evidence that there was a statistically significant difference in the participants' mean score for Narrative Prompt Response F and the FLDOE score.

Table 20 One-Sample *t*-Test: Writing Sample G (Post-Test)

	Test Value = 4					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Post-test narrative G	-1.768	16	.096	-.294	-.65	.06

Table 20 summarized the results of the *t*-test conducted to determine whether the study participants' mean score for Expository Prompt Response G differed from the score of 4 assigned by the Florida Department of Education. The test was conducted using an alpha of .05. The participants' mean score (M = 3.71, s = .686, SEM = .166) was not statistically significantly different from 4, t (16) = - 1.768, p > .05. The 95% confidence interval for the mean difference was -.65 to .06, and it did contain zero. In that case, the results provided evidence that there was not a statistically significant difference in the participants' mean score for Expository Prompt Response G and the FLDOE score.

Table 21 **One-Sample *t*-Test: Writing Sample H (Post-Test)**

	Test Value = 3					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Post-test narrative H	2.729	16	.015	.529	.12	.94

Table 21 summarized the results of the *t*-test conducted to determine whether the study participants' mean score for Narrative Prompt Response H differed from the score of 3 assigned by the Florida Department of Education. The test was conducted using an alpha of .05. The participants' mean score (M = 3.53, s = .800, SEM = .194) was statistically significantly different from 3, t (16) = 2.729, $p < .05$. The 95% confidence interval for the mean difference was .12 to .94, and it did not contain zero. In that case, the null hypothesis was rejected. The results provided evidence that there was a statistically significant difference in the participants' mean score for Narrative Prompt Response H and the FLDOE score.

In sum, there was a statistically significant difference between the participants' mean scores and the FLDOE scores in 2 out of 4 cases. There was significant agreement between the score of 5 assigned by the Florida Department of Education and the participants' mean score for Expository Prompt Response E. There was also significant agreement between the score of 4 assigned by the Florida Department of Education and the participants' mean score for Narrative Prompt Response G.

Intraclass Correlations

As in the pre-test, Cronbach's alpha was the statistic used to assess inter-rater reliability. Using a two-way mixed-model, it was found that for the four post-test scores, the alpha coefficient was .919, a high correlation (Cherry & Meyer, 1993). It is again noted that the high alpha coefficient should be interpreted with caution because when the rating scale is small, as in the 1-6 rating scale used in this study, the reliability coefficients tend to increase as the number of raters increase (Moore & Young, 1997; Cherry & Meyer, 1993).

Post-Test Percentage Agreement Analysis

An evaluation of the pre-test percentage agreement was conducted and is outlined below in Table 22. Table 22 shows percentages for the number of times the participants' scores were in complete agreement with or within one point of the FLDOE score for each of the four pre-test samples.

Table 22 Post-Test Percentage Agreement - Participants' Scores and FLDOE

	Number of scores	Number in exact agreement	Percentage in exact agreement	Number within 1 score point	Percentage within 1 score point
Expository Sample E	17	8	47.06	15	88.24
Expository Sample F	17	5	29.41	17	100.00
Narrative Sample G	17	8	47.06	17	100.00
Narrative Sample H	17	8	47.06	17	88.24
Total	68	29	42.65	64	94.12

On the post-test, scores assigned by the study participants were in exact agreement with the FLDOE in 29 out of 62 cases, or 42.65% of the time. They were within one point of the FLDOE in 64 out of 68 cases, or 94.12% of the time.

CHAPTER FIVE: CONCLUSION

This study sought to examine the benefits of a training activity that was planned, implemented, and evaluated at one school site. The training was an attempt to improve the inter-rater reliability of scores assigned by teachers to fourth-grade students' essays and stories. The assigned scores were based on the students' demonstration of writing skills as outlined on a state-prescribed 6-point holistic rubric (Appendix C). The training participants were 17 teachers assigned to teach in grades three through five. The two instructors were also members of the school faculty, but their backgrounds and experiences classified them as "staff experts" in writing instruction and assessment. The participants engaged in three training sessions. Session 1 was a basic lecture format, while Sessions 2 and 3 focused on practice scoring and group discussions. A pre-test was administered prior to the training, and a post-test was administered at its conclusion.

Additionally, a qualitative study of the relationship between teachers' writing confidence levels [as measured by the *Teacher/Writer Questionnaire* (Bowie, 1996)], the assignment of scores, and response to training was also conducted. Six participants, three with the highest levels of confidence as measured by the survey instrument and three with the lowest levels, were targeted for comparison.

This research study addressed the following questions:

- Will teachers benefit from in-house training designed to align teachers' evaluations of students' writings to standards established by the state?
- Will teachers' perceptions of their personal writing ability and their approach to writing tasks influence their assessment of students' writings?

Summary of the Findings

The summary is divided into two sections. The first section is limited to data collected prior to and after the training sessions. It is an examination of the general effectiveness of the training and addresses the first hypothesis:

- Exposing raters of student writings to directed discussion sessions and giving them the opportunity to identify exemplars of student performance at various levels of the performance scale will significantly improve the consistency of assigned scores and align them with standards set by the state.

The second section combines quantitative and qualitative measures. It examines the scoring behaviors and patterns of six of the study participants. The six targeted participants were selected because of their reported apprehensions, or lack of, related to writing and writing tasks. The three participants who reported the least writing apprehension were classified as the "high-confidence" group, and the three who reported the most writing apprehension were classified as the "low-confidence" group. The mean scores assigned by both groups were compared across groups and to the scores assigned by the Florida Department of Education (FLDOE). Additional comparisons included percentage agreement analyses and an examination of the connotations (positive or negative) of written comments. That section addresses the second hypothesis:

- Teachers with low levels of writing apprehension will tend to assign lower scores to students' writings than their "less-confident" peers.

Section 1: Pre-Test – Post-Test Comparisons

The 17 teachers scored four writing samples prior to the training and four writing samples after training. Several analyses were used to assess the consistency of those scores. The analyses included *t*-tests that compared the participants' mean scores to the scores assigned by the Florida Department of Education (FLDOE), a within-group analysis of reliability as measured by Cronbach's Alpha, and percentage agreement analyses.

T-Tests.

The pre-test writing responses were scored independently, without discussion or verbal exchanges by the participants. All participants were provided a full copy of the 6-point holistic rubric (Appendix C) to use during the scoring process. All scoring was done in one location. The writing samples selected for the pre-test and post-test were deemed to be representative of Levels 3, 4, and 5 by the FLDOE. They were selected from the *Florida Comprehensive Assessment Test 2006 and 2007 Writing+ Training Materials: Anchor Papers and Qualification Sets* (Florida Department of Education computer software, 2006 & 2007). Levels 1, 2, and 6 were not used for the pre-test and post-test because Levels 1 and 2 were the minimal levels, and Level 6 was the highest. Since papers at those three levels were at the extreme ends of the spectrum, they were considered less likely to generate scores both above and below the FLDOE score. For example, a paper scored at a Level 6 by the FLDOE was likely to be evaluated as a "high-end" paper by all participants, but the score assigned by the participants could not be higher than Level 6. Similarly, the lower levels were less likely to generate disagreement among raters. According to the FCAT rubric, a Level 1 response minimally addresses the topic and does not exhibit an organizational pattern. A Level 2 response is similarly immature and may contain little relevant

or supporting information. Level 1 and Level 2 responses provided by the Florida Department of Education (2006 & 2007) in its training software packages, *Florida Comprehensive Assessment Test 2006 and 2007 Writing+ Training Materials: Anchor Papers and Qualification Sets*, were significantly shorter than responses at the other 4 levels: most of the Level 1 and 2 responses consisted of less than 10 lines of writing. The middle level responses, Levels 3, 4, and 5 were lengthier and therefore likely to require more thoughtful evaluation by the raters. They were also the levels the teachers at the school reportedly had the most difficulty evaluating.

There were four writing samples, two expository and two narratives used in the pre-test. According to the FLDOE raters, Expository Response A was Level 4, Expository Response B was Level 3, Narrative Response C was Level 5, and Narrative Response D was Level 4. The *t*-test results showed that the mean scores of the study participants were statistically significantly different from the scores assigned by the Florida Department of Education in three out of four cases. The participants' assessment of Expository Prompt Response B was the most similar to the FLDOE score.

Similarly, there were four writing samples, two expository and two narratives used in the post-test. According to the FLDOE raters, Expository Response E was Level 5, Expository Response F was Level 4, Narrative Response G was Level 4, and Narrative response H was Level 3. The *t*-test results showed that the mean scores for the study participants were statistically significantly different from the scores assigned by the FLDOE in two out of four cases. The participant's assessment of Expository Response E and Narrative Response G were the most similar to the FLDOE scores.

Considering the *t*-tests alone, it appeared that there was only a slight improvement in the participants' ability to match the FLDOE score: from 1 out of 4 cases on the pre-test, to 2 out of

4 cases on the post-test. Of the three instances where the participants' mean score were similar to the FLDOE scores (Responses B, E, and G), there seemed to be no particular relationship to any particular score level as each response, B, E, and G, represented a different level.

Within Group Comparisons.

There was little or no difference in the reliability coefficient (Cronbach's Alpha) produced by the pre-test scores (.913) and the reliability coefficient produced by the post-test scores (.919). While there seemed to be a considerable amount of discrepancy between the mean scores of the participants and the scores assigned by the FLDOE as noted in the *t*-test results, the alpha coefficients reflect a picture of within-group reliability both prior to and after the training. Several researchers, however, have noted that the alpha coefficient is less than an ideal measurement when comparing multiple raters (Moore & Young, 1997; Cherry & Meyer, 1993). Cherry and Meyer (1993) pointed out that during training, it is often the practice to have a small number of writing responses scored by a large number of raters. While it is interesting to note that some level of scoring consistency existed among the 17 raters both prior to and after the training, no conclusions can be based solely on the alpha coefficients. The problem is that the reliability measures produced under those circumstances will not be the same as the reliability measures produced under more practical circumstances, when only one or two raters score a given text. Under normal circumstance at the school in the study, student responses are seldom read by more than two raters, and in most cases, they are only read by one. In light of the *t*-test results and the percentage agreement analyses that follow, it would be reasonable to believe that the alpha coefficients, in the case of this study, projected a measure of reliability that was somewhat inflated.

Percentage Agreement Analyses.

The percentage agreement analyses presented the clearest picture of the effect of the in-house training. Prior to the training, the seventeen participants matched the FLDOE score in 20 out of 68 cases, or 29.41% of the time. They matched or were within one score point of the FLDOE in 54 of 68 cases, or 79.41% of the time. After the training, participants matched the FLDOE scores in 29 out of 62 cases, or 42.65% of the time. They matched or were within one score point of the FLDOE in 64 out of 68 cases or 94.12 % of the time. In sum, from pre-test to post-test, there was a 13% increase in the number of matched scores and a 15% increase in the number of scores within one score point of the FLDOE.

Conclusions about the Effects of the Training.

This study was small, so any conclusions are necessarily tentative. However, the results suggested that the in-house training activities promoted higher inter-rater reliability of scores assigned to students' writings by the teachers in this study. The slight improvement in agreement noted by the *t*-tests, from 1 in 4 cases to 2 in 4 cases, combined with the much higher post-test percentage agreements indicated that there was some evidence of improved inter-rater reliability due to training. As noted, a much tighter range of scores was evident on the post-test evaluations as opposed to the pre-test evaluations. For example, post-test scores were within one point of the FLDOE score 94% of the time, a considerable improvement over the pre-test where scores were within one point only 79% of the time. While the within group reliability coefficient (Cronbach's Alpha), was relatively the same both prior to and after the training, the percentage agreement analyses and *t*-test results indicated that there was a significant level of difference between the

pre-test scores assigned by the study participants and the scores assigned by the FLDOE. The data supports the hypothesis that exposing raters to directed discussion sessions and giving them the opportunity to identify exemplars of student performance at various levels of the performance scale will improve the consistency of assigned scores and align them with standards set by the state.

Section 2 – A Comparison of "High-Confidence" and "Low-Confidence" Groups

A writing apprehension score and rank order (Appendix B) was based on the participants' responses to 12 of the statements on the Teacher/Writer Questionnaire (Bowie, 1996). The three participants with the lowest apprehension scores and the three participants with the highest apprehension scores were identified and targeted for further comparisons. After the rankings were determined, it was noted that there were no significant patterns related to education or experience. A summary of the more noteworthy comparisons follow.

Mean Comparisons.

The results suggested that before the training, teachers who were "highly-confident" about their own abilities as writers were more likely to assign scores that were lower than the FLDOE and lower than the scores assigned by their "less-confident" peers. On the pre-test, the mean scores of the "high-confidence" group were lower than the FLDOE scores in 4 out of 4 cases, 100% of the time. The mean scores of the "high-confidence" group were lower than the mean scores of the "low-confidence" group in 3 out 4 cases, 75% of the time. There was one case in which the mean score of the "high-confidence" group was equal to the mean score of the "low-confidence" group. In other words, on the pre-training assessments, there were no cases

where the mean scores of the "high-confidence" group were higher than the scores of the "low-confidence" group or the FLDOE.

On the post-training assessments, the score patterns were balanced. The mean scores of the "high-confidence" group were lower than the FLDOE scores in 2 out of 4 cases and higher than the FLDOE scores in 2 out of 4 cases, both 50% of the time. Cross group comparisons produced the same pattern. The "high-confidence" groups' mean scores were lower than the mean scores of the "low-confidence" group in 2 out of 4 cases and higher than the mean scores of the "low-confidence" group in 2 out of 4 cases, both 50% of the time.

Percentage Agreement.

The results suggested that the training improved the reliability of scores assigned by both groups of participants, but the "less-confident" teachers displayed more consistency on both the pre-test and the post-test than the teachers with higher confidence levels. On the pretest, both groups were in exact agreement with the FLDOE 42% of the time. The "high-confidence" group's scores were within one point of the FLDOE score 67% of the time, while the "low-confidence" group's scores were within one point of the FLDOE 92% of the time. On the post-test, the "high-confidence" groups' scores were equal to the FLDOE score 33% of the time and were within one score point of the FLDOE score 83% of the time. The "low-confidence" group's scores were in exact agreement with the FLDOE 50% of the time and within one score point 100% of the time.

Extended Comparisons.

In addition to the pre-test and post-test scoring exercises, individual scoring took place during two of the training sessions, Session 2 and Session 3. During those sessions a total of 14 writing samples were rated. Unlike the pre-test and post-test, those samples spanned all six of the score levels. The scoring patterns and behaviors of the "high-confidence" and "low-confidence" groups are outlined in the sections that follow.

Inter-Rater Reliability.

The reliability coefficient (Cronbach's Alpha) was high for both groups. The coefficient for the "high-confidence" group was .886. The coefficient for the "low-confidence" group was .936. While the "low-confidence" group displayed a greater level of internal consistency, intraclass coefficients are considered high if they are above .80 (Cherry & Meyer, 1993).

Within-group percentage agreement analyses were also conducted on the scores assigned during Sessions 2 and 3. In Session 2, the "high-confidence" group slightly outperformed the "low-confidence" group. The "high-confidence" participants' scores differed by less than one score point on 6 of the 7 writing samples assessed during that session. The "low-confidence" participants' scores differed by less than one score point on 5 of the 7 writing samples.

In Session 3, the "low-confidence" group outperformed the "high-confidence" group. The "highly-confident" participants' scores differed by less than one score point on only 3 of the 7 writing samples assessed during that session, while the "less-confident" participants' scores differed by less than one score point on 7 of the 7 writing samples. This pattern seemed to

suggest that as the training progressed, the inter-rater reliability of the members of the "low-confidence" group improved significantly.

Group Influences.

The examination of how frequently the members of the two groups changed their scores as a result of discussion revealed that there was little or no differences between groups. In addition, members of both groups were just as likely to increase or decrease their scores after discussion with other participants.

Comment Analyses.

During Sessions 2 and 3, participants were asked to independently assign scores to 14 writing samples and to give rationales for their decisions. Each comment was identified as being either positive or negative in nature. For example, a comment such as "attempts organization" was classified as a positive comment, while "lacks organization" was classified as a negative comment. The negative comments made by the members of the two groups were similar in number: 26 in the "high-confidence" group; 23 in the "low-confidence" group. However, the number of positive comments made by the "low-confidence" group (74) was almost double the number of positive comments made by the "high-confidence" group (39). In sum, it can be noted that neither group was excessively negative when stating rationales for the scores assigned during Sessions 2 and 3. While members of both groups were more likely to write positively phrased comments rather than negative comments, the members of the "low-confidence" group made a significantly higher number of positive comments than the "high-confidence" group.

Outliers.

As the group comparisons were conducted, it was noted that two individuals, Alice, the participant with the highest confidence rank, and Quincy, the individual with the lowest confidence rank, accounted for some of the more pronounced differences between groups. For example, Alice accounted for more than 50% of the individual scores that differed from the group consensus in both Sessions 2 and 3. In contrast, Quincy accounted for 33% of the scores that differed from the group consensus in Session 2 and 0% of the scores that differed from the group consensus in Session 3. Another contrast presented by the responses of those two participants was noted in the comment analyses. Alice made the smallest number of positive comments (6) while Quincy made the largest number of positive comments (41). While no clear conclusions can be dawn from that data, it is somewhat unusual and might indicate a need for further research. It might be speculated that compared to Quincy, the "highly confident" Alice was less receptive to the training because her scoring patterns showed that her individual scores were just as likely to match as not match her group's consensus scores. On the other hand, the individual scores assigned by the "less-confident" Quincy frequently matched his group's consensus scores. In addition, the large number of comments Quincy wrote in comparison to Alice might indicate that he was more thorough, and possibly more fully engaged in the assessment process.

Conclusions: Influences of Personal Backgrounds and Experiences.

The data related to the performances of the "high-confidence" and "low-confidence" groups in this study does somewhat support the hypothesis that "highly-confident" teachers tend to assign lower scores to students' writings than their "less-confident" peers. However, in this

study, that tendency seemed limited to the pre-training assessments. Those assessments indicated that the members of the "high-confidence" group were initially more critical than the members of "low-confidence" group, but subsequent analyses of scoring patterns showed little or no differences between the value (high or low) of the scores assigned by the two groups. An example of the flexibility of the "high-confidence" group is shown in the data collected during Sessions 2 and 3. During group discussions held during those sessions, members of the "high-confidence" group were just as likely to change their scores to a higher level as to a lower level. In sum, the percentage agreement analyses indicated that both the "high-confidence" group and the "low-confidence" group benefited from the training because both groups demonstrated more consistent and more balanced score patterns on the post-training assessments.

Some data does indicate that the "low-confidence" group was more receptive to the training than the "high-confidence" group. For example, on the post-test, the "low-confidence" group's scores were equal to or within one score point of the FLDOE score 100% of the time, while the "high-confidence" group's percentage agreement on the same assessment was only 83%. Members of the "low-confidence" group were also less likely to omit a rationale for the scores they assigned. Since all data was collected anonymously, it is unclear why the "high confidence" group frequently failed to note rationales. One might speculate that the papers without rationales were more obviously within a particular score level than others, or that the "high-confidence" group did not feel the need to justify the scores they assigned. Further research would be needed before any attempt could be made to address those issues.

Discussion

This study took shape because there was an identified need at the school in the study which may well be an issue at similar schools. A significant turnover in staff combined with grade reassignments created a need for teachers to "get on the same page." The school administration and teachers expressed a desire to strengthen their knowledge of writing assessment and thereby clear a path for aligning instruction to state standards. This discussion will summarize the key issues and findings of this study by reviewing the challenges associated with writing assessment, the importance of establishing realistic goals and expectations, and the effect teachers' beliefs in their personal efficacy as writers might have on the assessment process.

The Challenge of Writing Assessment

Unlike other subject areas, writing instruction and assessment is less structured and therefore less likely to fit into a neat package (White, 1993). Trimbur (1996) explained that unlike the teaching of literature, writing instruction and testing is more complicated. "After all, we don't just expose students to writing; we expect them to acquire some demonstrative skill at it, and we have spent a lot of time and effort figuring out how to measure this skill" (p. 47).

In the early decades of large-scale standardized testing, it was the norm to assess writing through "objective" multiple-choice tests (Bracey, 2002; Yancey; 1999). In this age of accountability, high-stakes testing is viewed as a means of reform (Linn, 2000). Polls show that the public overwhelming supports the use of standardized tests (Phelps, 2005), and many states have attempted to align their testing programs to their state standards (Higgins, Miller, & Wegmann, 2007). States have gradually moved away from the indirect assessment of writing

because such superficial assessment of writing skills has little relationship to real-world applications (Lyman, 1998). The simple and logical argument in favor of direct writing assessment is that a test of writing skills should involve writing (Bracey, 2002; Diederich, 1974).

Florida is one state that has attempted to embrace both direct and indirect writing assessment. Since 2005, it has administered the Florida Comprehensive Assessment Test (FCAT) Writing + to fourth, eighth, and tenth-grade students (Florida Department of Education, 2008). The direct writing component of the FCAT Writing + is scored according to a 6-point holistic rubric. Many schools and districts recommend interim tests that follow the same format and use the same scoring rubric as the state mandated assessments (Orange County Public Schools, 2008; Seminole County Public Schools, 2007).

This study used the FCAT Writing + holistic rubric as its scoring guide. The in-house training was an attempt to improve the consistency of scores teachers at the school in the study assigned to the writings of upper elementary school students. The training consisted of three sessions, two of which focused on scoring practice and discussion. The format was consistent with the recommendations outlined by Maki (2004) and similar to the rater training conducted by Myerberg (1996). All data was collected anonymously, and the training participants reported feeling comfortable with the training process.

The participants in this study were faced with the challenge of learning to interpret the holistic rubric in a similar manner and thereby ensure a level of consistency in scoring students' writings. The 6-point holistic writing rubric used to score the FCAT Writing + expository and narrative writing responses did contain what some of the participants considered vague language. Nonspecific phrases such as "generally correct," "generally followed," "generally focused," and "generally adequate" were liberally used as descriptors (See the FCAT holistic rubric in

Appendix C). Bainer and Porter (1992) conducted a study that required teachers to use a holistic rubric and reported similar concerns:

> Most teachers expressed difficulty understanding and discriminating between some levels of the rubric. Some reported that they would "sit and agonize whether it was a 2 or 3, concluding that in the end it was 'strictly a judgment call'" and that "the more you read (the papers), the more you change your mind" (p. 13).

White (1986) stressed the need for professional respect and community among raters that would allow them to respond similarly to the texts they evaluated. The presenters who directed the training in this study attempted to clarify what some participants called the "mysterious" rubric, but the bigger challenge was creating a "community" as White described. The two sessions that allowed the participants to share their ideas and rationales as they participated in the scoring process were developed in the hope of achieving that goal.

The results indicated that some level of success was achieved. On the post-training assessments the study participants narrowed the range of scores assigned to prompt responses. The scores the participants assigned to the post-training assessments were also more closely matched to the scores of the Florida Department of Education. This finding is consistent with Maki's (2004) claim that improved score reliability can be achieved with as few as two scoring/discussion sessions.

If the purpose of district and school level writing assessments is to evaluate performance, provide feedback, and inform instruction, it seems fitting that there be some in-house training regarding performance standards. The question arises as to how schools and districts can improve the reliability of holistically scored writing responses within a limited time and with

limited resources. The results of this study indicated that there are practical and cost-efficient means of providing school faculty members support and opportunities for growth.

Goals and Expectations

When it comes to writing assessment, some level of disagreement among raters is acceptable. Smith (1993) pointed out that essays are not as neat and tidy as the rubrics used to assess them and that we need to guard against the assumption that disagreement among raters necessarily equates to lack of reliability. The goal of the training presented as part of this study was not to have *all* the raters agree *all* the time but to lower the point range of the disagreements. Studies have shown that it is much more difficult to consistently assess writing than other performance tasks (Myerberg, 1996; Van Noord & Prevatt, 2002). White (1993) attested that some writings resist agreement and that there is nothing wrong with reporting the average of two scores if those scores are within one point of each other. In the last three years, 38 - 41 percent of the scores assigned to fourth-grade essays on the FCAT Writing + was an average of two scores (Florida Department of Education FCAT Writing Scores [Data file], 2007). That is, if one rater assigned a score of 3 to a particular paper and the second rater assigned a score of 4, the average score of 3.5 was reported to parents and schools.

Teachers might feel more comfortable about being the sole scorer of an essay or story if allowed to augment scores when they perceive that writing responses "straddle the fence." Like the training participants in this study, Smith (1993) pointed out that there are "gray" areas associated with holistic rubrics, and we cannot assume that a holistic scale adequately covers the performance range for written language. In his research of placement test essay evaluators, Smith (1993) noted that raters often "put little pluses or minuses next to their ratings, presumably

because they are uncomfortable with the scale" (p. 196). As Penny, Johnson, and Gordon (2000) found, when allowed to augment scores with a (+) or minus (-), raters chose to do so almost 50% of the time and inter-rater reliability improved significantly. When only one rater is used, score augmentation might be a valid consideration at the school level.

Continuing training and monitoring within-school writing assessments would seem appropriate if such assessments are to be used to effectively monitor progress and inform instruction. Training sessions conducted by school faculty may be a very cost-effective way of developing community and ensuring fairness. As White (1993) put it, "Reliability is a technical way of talking about simple fairness to test takers, and if we are not interested in fairness, we have no business giving tests or using test results" (p. 93). Given that members of school faculties have varied backgrounds and years of experience, it is important that scoring rubrics used to assess students' writings are clearly understood and that there is a general consensus among faculty members as to what constitutes effective writing.

Self-Efficacy

It was obvious from the participants' responses to questions on the *Teacher/Writer Questionnaire* (Bowie, 1996) that varying levels of confidence existed among the participants in regards to their personal writing abilities (Appendix B). The comparison study of the three participants with the highest reported confidence levels and the three with the lowest reported confidence levels provided additional insights. The researcher expected the highly confident teachers to be "tougher" scorers than their less confident peers. As Faigley, Daly, and Witte (2001) reported, individuals with high levels of writing apprehension tend to avoid writing tasks. This study's researcher felt that lack of knowledge and limited positive experiences would make

the less confident participants more likely to assign inflated scores. The results showed that prior to the training, the highly confident teachers did tend to assign lower scores to students' writings than their less confident peers, but after training, there were no significant differences. This study showed that the training may have had its desired effect in creating a team of raters who were more objective and hopefully better informed than they were prior to the training.

Routman (2000) noted that elementary teachers receive very limited training on how to teach writing. While not a focus of this particular research, it is appropriate to note that an overwhelming number of this study's participants (76%) disagreed or strongly disagreed with the statement: "My teacher education program has trained me well to teach writing." Even more dramatic was the fact that 88% agreed or strongly agreed with the statement: "Writing assignments are difficult to grade" (Appendix A). Those responses underscore the need for continued training in the area or writing and writing assessment at the school in this study.

Recommendations

This study involved a small group of teachers on staff at one elementary school; therefore, there remain unanswered questions about writing evaluation, training, and the influence of teachers' self-efficacy that warrant consideration. Recommendations for further research follow.

- Further study of the impact of in-house training as outlined in this study could be conducted at other schools and in larger numbers to determine whether the results are replicable.

- A similar study could be conducted to determine whether varying the number of weeks between sessions significantly impacts the results.
- The study could be expanded to include follow-ups to determine whether improved inter-rater reliability rates are maintained over time.
- Some version of the training sessions outlined in this study should be repeated annually for new or reassigned staff members at the school in this study.
- A study could be conducted to determine whether writings scored by language arts teachers at other school levels are more or less likely to be aligned to state standards.
- A larger study of elementary school teachers' attitudes and beliefs about writing and writing tasks and how those beliefs influence instruction and assessment would seem appropriate.

Holistic rubrics used to assess written language are likely to remain a part of our educational testing environments for some time (Goodman & Hambleton, 2005). The challenge for districts and schools who adopt on-demand writing assessments as part of local assessment programs is to provide fair evaluations. This study is one example of how schools might conduct site-based training activities that improve scoring consistency and ensure that the writing assessments they administer are valid and reliable.

APPENDIX A: RESPONSES TO BOWIE'S (1996) TEACHER/WRITER QUESTIONNAIRE

Questionnaire Statements	Frequencies					
	SA	A	U	D	SD	Mean
1. I avoid writing.	0	2	1	8	6	4.06
2. I have no fear of my writing being evaluated.	1	6	4	4	2	3.00
3. I look forward to writing down my ideas.	4	9	3	4	0	2.06
4. Teachers in my field do not have to be writers.	0	3	2	8	4	3.76
5. I'm afraid of writing essays when I know they will be evaluated.	1	5	3	5	3	3.24
6. Teachers should write along with their students.	8	7	2	0	0	1.65
7. Handing in a composition makes me feel good.	2	7	6	1	1	2.53
8. My mind seems to go blank when I start work on a composition.	1	2	4	8	2	3.47
9. Expressing my ideas through writing seems to be a waste of time.	0	0	2	10	5	4.18
10. Writing assignments are difficult to grade.	5	10	0	2	0	1.94
11. I would enjoy submitting my writing for evaluation and publication.	0	4	6	5	2	3.29
12. I like to write down my ideas.	3	11	2	1	0	2.06
13. I feel confident in critiquing another person's writing.	0	7	1	7	2	3.24
14. Writing should be incorporated in all classes.	6	9	2	0	0	1.76
15. I feel confident in my ability to clearly express my ideas in writing.	1	11	3	1	1	2.41
16. I like to have my friends read what I have written.	1	3	5	6	2	3.29
17. I'm nervous about writing.	1	5	2	6	3	3.29
18. Writing is more important in some classes than others.	1	6	2	8	0	3.00
19. People seem to enjoy what I write.	1	6	8	1	1	2.71
20. I do not need instruction in writing.	0	2	4	8	3	3.71

Questionnaire Statements	Frequencies					
	SA	A	U	D	SD	Mean
21. I enjoy writing.	4	7	2	3	1	2.41
22. I never seem to be able to clearly write down my ideas.	0	1	3	11	2	2.82
23. Writing is a lot of fun.	3	8	3	2	1	2.41
24. All teachers should be writers.	1	9	3	3	1	2.65
25. I expect to do poorly in composition classes even before I enter them.	0	1	3	6	7	4.12
26. I like seeing my thoughts on paper.	2	8	5	1	1	2.47
27. I plan to use writing regularly in my classes when I teach.	4	10	2	1	0	2.00
28. Discussing my writing with others is enjoyable.	3	3	3	6	2	3.06
29. I have a terrible time organizing my ideas when writing.	0	3	4	7	3	3.59
30. When I hand in a composition, I know I am going to do poorly.	0	0	2	8	7	4.29
31. Mathematics does not lend itself well to writing.	0	1	1	10	5	4.12
32. It's easy for me to write good compositions	1	7	5	3	1	2.76
33. When teaching, I try to correct all my students' writing mistakes.	0	6	0	7	4	3.53
34. I don't think I write as well as most other people.	1	6	2	7	1	3.06
35. I don't like my composition to be evaluated.	2	8	2	4	1	2.65
36. Whether or not I write has no bearing on my students' writing.	0	1	0	12	4	4.12
37. I'm not good at writing.	0	4	2	9	2	3.53
38. I want to teach writing.	4	8	0	2	3	2.53
39. Taking a composition course is a very frightening experience.	0	4	2	8	3	3.59
40. My teacher education program has trained me well to teach writing.	0	2	2	6	7	4.06

APPENDIX B: SELF-CONFIDENCE RANKINGS

	Survey Questions													
	q2	q7	q11	q15	q20	q22	q30	q32	q34	q35	q37	q39	Total	Rank
Alice	2	1	2	1	4	1	1	1	1	1	1	1	17	1
Betty	1	2	2	2	2	2	2	2	2	2	2	2	23	2
Calvin	2	2	3	2	4	1	1	2	2	2	1	2	24	3
Diane	4	2	2	2	4	2	1	2	2	2	2	1	26	4
Ellen	2	2	4	2	2	2	1	2	2	4	2	2	27	5
Frank	2	3	4	2	3	2	1	2	2	2	2	2	27	5
Gerald	2	1	3	2	4	2	1	2	4	4	2	1	28	6
Helen	3	2	3	2	3	2	2	2	3	3	2	2	29	7
Ivan	4	2	2	2	4	2	2	4	2	4	2	2	32	8
James	2	4	3	2	3	2	2	3	4	4	3	2	34	9
Kathy	5	2	4	2	5	2	2	3	2	4	2	2	35	10
Louise	3	3	3	3	4	3	2	3	3	3	3	3	36	11
Monica	3	3	3	2	5	2	2	3	4	4	2	4	37	12
Nora	3	5	5	3	3	3	3	3	4	4	4	3	43	13
Opal	4	3	4	3	4	3	2	4	4	5	4	4	44	14
Patrick	4	3	4	4	4	2	3	4	4	4	4	4	44	14
Quincy	5	3	5	5	5	4	1	5	5	5	4	4	51	15

APPENDIX C: FCAT WRITING RUBRIC (FLORIDA DEPARTMENT OF EDUCATION FCAT HANDBOOK, 2005)

6 Points: The writing is focused on the topic, has a logical organizational pattern (including a beginning, middle, conclusion, and transitional devices), and has ample development of the supporting ideas. The paper demonstrates a sense of completeness or wholeness. The writing demonstrates a mature command of language including precision in word choice. Subject/verb agreement and verb and noun forms are generally correct. With few exceptions, the sentences are complete, except when fragments are used purposefully. Various sentence structures are used.

5 Points: The writing is focused on the topic with adequate development of the supporting ideas. There is an organizational pattern, although a few lapses may occur. The paper demonstrates a sense of completeness or wholeness. Word choice is adequate but may lack precision. Most sentences are complete, although a few fragments may occur. There may be occasional errors in subject/verb agreement and in standard forms of verbs and nouns, but not enough to impede communication. The conventions of punctuation, capitalization, and spelling are generally followed. Various sentence structures are used.

4 Points: The writing is generally focused on the topic, although it may contain some extraneous or loosely related information. An organizational pattern is evident, although lapses may occur. The paper demonstrates a sense of completeness or wholeness. In some areas of the response, the supporting ideas may contain specifics and details, while in other areas, the supporting ideas may not be developed. Word choice is generally adequate. Knowledge of the conventions of punctuation and capitalization is demonstrated, and commonly used words are usually spelled correctly. There has been an attempt to use a variety of sentence structures, although most are simple constructions.

3 Points: The writing is generally focused on the topic, although it may contain some extraneous or loosely related information. Although an organizational pattern has been attempted and some transitional devices have been used, lapses may occur. The paper may lack a sense of completeness or wholeness. Some of the supporting ideas may not be developed with specifics and details. Word choice is adequate but limited, predictable, and occasionally vague. Knowledge of the conventions of punctuation and capitalization is demonstrated, and commonly used words are usually spelled correctly. There has been an attempt to use a variety of sentence structures, although most are simple constructions.

2 Points: The writing may be slightly related to the topic or may offer little relevant information and few supporting ideas or examples. The writing that is relevant to the topic exhibits little evidence of an organizational pattern or use of transitional devices. Development of the supporting ideas may be inadequate or illogical. Word choice may be limited or immature. Frequent errors may occur in basic punctuation and capitalization, and commonly used words may frequently be misspelled. The sentence structure may be limited to simple constructions.

1 Point: The writing may only minimally address the topic because there is little, if any, development of supporting ideas, and unrelated information may be included. The writing that is relevant to the topic does not exhibit an organizational pattern; few, if any, transitional devices are used to signal movement in the text. Supporting ideas may be sparse, and they are usually provided through lists, clichés, and limited or immature word choice. Frequent errors in spelling, capitalization, punctuation, and sentence structure may impede communication. The sentence structure may be limited to simple constructions.

APPENDIX D: INDIVIDUAL SCORE SHEET

Sample #	Initial Score	Rationale	Group Score	Rationale	DOE Score
A					
B					
C					
D					
E					
F					
G					

APPENDIX E: SUMMARY OF RATIONALES FOR SCORES ASSIGNED TO WRITING SAMPLE F4

ID	Ind. Score	Individual Rationale	Group Score	Group Rationale	FLDOE Score = 4 Rationale
Alice	3	Repetitions	3	Formula Writing Style	The response is focused on the topic, and an organizational pattern is evident. Support consists of three elaborated ideas with some specific details. Word choice is adequate, and knowledge of basic conventions is demonstrated. (Florida Department of Education, 2006).
Betty	4	Good O and Focus	4	EL, good transition	
Calvin	3		4	F, O, EL,	
Opal	5	Transitional words; EX-told specific instances	4	Elaborations	
Patrick	4	Transitional words, sense of completeness	4	EL	
Quincy	3		3	Focus, O, Ela spec. details	

APPENDIX F: SUMMARY OF RATIONALES FOR SCORES ASSIGNED TO WRITING SAMPLE I9

ID	Ind. Score	Individual Rationale	Group Score	Group Rationale	FLDOE Score = 2 Rationale
Alice	3	Conv imp com + lapses	2	Conventions impeded comprehension	The response is focused on the topic, and a simple story line is attempted. Support consists of bare and extended events. Errors in conventions occur, but do not impede communication (Florida Department of Education, 2006).
Betty	2	BI, Lacks EL & EX	2	Word choice is simple	
Calvin	3	BI	3	BI	
Opal	3	Organization-sequence of time; transitional words not always used correctly	3	BI	
Patrick	2	Errors in common words; Errors impeded comprehension; Beginning was confusing	2	Conv. impeded comprehension; lapses	
Quincy	2	F, O, C, BI	2	Word choice is simple	

APPENDIX G: UNIVERSITY OF CENTRAL FLORIDA INSTITUTIONAL REVIEW BOARD APPROVAL NOTICE

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Notice of Expedited Initial Review and Approval

From : **UCF Institutional Review Board**
FWA00000351, Exp. 5/07/10, IRB00001138

To : **Lisa Farmer**

Date : **October 31, 2007**

IRB Number: **SBE-07-05287**

Study Title: **Enhancing Reliability of Teachers' Holistic Scoring of Elementary Writing Through In-House Professional Development**

Dear Researcher:

Your research protocol noted above was approved by **expedited** review by the UCF IRB Vice-chair on 10/30/2007. **The expiration date is 10/29/2008.** Your study was determined to be minimal risk for human subjects and expeditable per federal regulations, 45 CFR 46.110. The category for which this study qualifies as expeditable research is as follows:

6. Collection of data from voice, video, digital, or image recordings made for research purposes.

7. Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

The IRB has approved **a consent procedure which requires participants to sign consent forms.** Use of the approved, stamped consent document(s) is required. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Subjects or their representatives must receive a copy of the consent form(s).

All data, which may include signed consent form documents, must be retained in a locked file cabinet for a minimum of three years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained on a password-protected computer if electronic information is used. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

To continue this research beyond the expiration date, a Continuing Review Form must be submitted 2 – 4 weeks prior to the expiration date. Advise the IRB if you receive a subpoena for the release of this information, or if a breach of confidentiality occurs. Also report any unanticipated problems or serious adverse events (within 5 working days). Do not make changes to the protocol methodology or consent form before obtaining IRB approval. Changes can be submitted for IRB review using the Addendum/Modification Request Form. An Addendum/Modification Request Form cannot be used to extend the approval period of a study. All forms may be completed and submitted online at http://iris.research.ucf.edu .

Failure to provide a continuing review report could lead to study suspension, a loss of funding and/or publication possibilities, or reporting of noncompliance to sponsors or funding agencies. The IRB maintains the authority under 45 CFR 46.110(e) to observe or have a third party observe the consent process and the research.

On behalf of Tracy Dietz, Ph.D., UCF IRB Chair, this letter is signed by:

Signature applied by Janice Turchin on 10/31/2007 10:34:42 AM EST

Janice M Turchin

IRB Coordinator

APPENDIX H: PARTICIPANT CONSENT FORM

Appendix G: Participant Consent Form

Third, fourth and fifth grade teachers at John Evans Elementary will participate in a special training program designed to improve the inter-rater reliability of teachers' holistic scores of student writings through in-house professional development. The training will be conducted in November and December of 2007, and will consist of three sessions, each approximately one hour long. Lisa Epps Farmer is conducting a research study in conjunction with that professional development in-service. While the training is mandatory for third, fourth, and fifth grade teachers on staff a John Evans Elementary, supplying data for the study is optional. The results of the study will be reported to Seminole County Public Schools and in a doctoral dissertation to be presented at the University of Central Florida some time in the near future.

A few days prior to the training, participants in the study will be asked to respond to a short questionnaire designed to gather information about teachers' attitudes and feelings related to writing tasks and assessments. At the beginning of Session 1, and at the end of Session 3, all third, fourth, and fifth grade teachers will be asked to score four student writings, two expository and two narrative responses. During Sessions 2 and 3, all third, fourth, and fifth grade teachers will be asked to score several student writings and participate in small group discussions related to those assigned scores.

At the conclusion of the training, the researcher will interview four participants. These interviews will be voluntary, and should last no more than 15 minutes. They will not be audio taped or videotaped. The interviewees will be asked four simple questions related to the effectiveness of the training. During the interviews, the interviewees may decline to answer any or all of the questions.

In order to keep participant's responses to the survey and the scores they assign student writings as confidential as possible, the researcher asks that each participating teacher assign himself or herself a four to six-character code name/number. That code will allow the researcher to associate the various pieces of data collected. The researcher will not keep a record of teacher codes, so participants are asked to supply a "hint" question or statement that might be used jog or refresh memory in the event a code is temporarily forgotten.

This research study has been reviewed and approved by the UCF Institutional Review Board. Questions or concerns about research participants' rights may be directed to the UCF IRB Office, University of Central Florida, Office of Research and Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL32826-16-3246. The telephone number is (407) 823-2901.

The researcher greatly appreciates your cooperation and hopes that the training will help teachers build consensus and gain insights into the writing needs of the Evans Elementary student population. If you have any questions or concerns, please feel free to contact Lisa (407-320-9834) or her UCF faculty advisor, Dr. Michael Hynes (407-823-2005).

I have read the procedure described above. I voluntarily agree to participate in the study conducted by Lisa Epps Farmer, and I have received a copy of this description. I understand that I may withdraw this consent at any time.

☐ By checking this box, I wish to express my willingness to participate in a short face-to-face interview at the end of the training.

Participant's Name (Please print.)_____

Participant Signature: _____ Date: _____

"Hint" question to be used in the event the participant forgets his/her code:
_____

University of Central Florida IRB
IRB NUMBER: SBE-07-05287
IRB APPROVAL DATE: 10/30/2007
IRB EXPIRATION DATE: 10/29/2008

REFERENCES

Abrams, L. M., Pedulla, J. J. & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice, 42* (1), 18-30.

Atkinson, D. & Murray, M. (1987, March). *Improving interrater reliability*. Paper presented at the Annual Meeting of the Conference on College Composition and Communication, Atlanta, GA.

Bainer, D. L. & Porter, F. (1992, October). *Teacher concerns with the implementation of holistic scoring.* Paper presented at the Annual Meeting of the Midwestern Educational Research Association, Chicago.

Baines, L. A. & Stanley, G. K. (2004). High-stakes hustle: Public schools and the new billion dollar accountability. *The Educational Forum (69)* 8-15.

Baldwin, D. (2004). A guide to standardized writing assessment. *Educational Leadership 62* (2), 72-75.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Best, J. W. & Kahn, J. V. (2003). *Research in education* (9th ed.). Boston: Pearson Education, Inc.

Bowie, R. L. (1996, October). *Future teachers' perceptions of themselves as writers and teachers of writing: Implications for teacher education programs.* Paper presented at the

Annual Meeting of the College Reading Association, Charleston, SC.

Bracey, G. W. (2002). *Put to the test: An educator's and consumer's guide to standardized testing.* Bloomington, ID: Center for Professional Development Phi Delta Kappa International.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M. & Rock, D. A. (1987). *Assessing writing skill.* New York: College Entrance Examination Board.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 1-6) Westport, CT: Praeger Publishers.

Brindley, R. & Schneider, J. J. (2002). Writing instruction or destruction: Lessons to be learned from fourth-grade teachers' perceptions on teaching writing. *Journal of Teacher Education, 53,* 328-341.

Brossell, G. (1986). Current research and unanswered questions in writing assessment. In K. L. Greenberg, H. S Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 44-52), New York: Longman.

Brown, R. (1986). A personal statement on writing assessment and education policy. In K. L. Greenberg, H. S Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 44-52), New York: Longman.

Bruner, J. S. (1971). *The relevance of education*. New York: W. W. Norton & Company, Inc.

Camp, R. (1993). Changing the model for writing assessment. In M. M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45-78), Cresskill, NJ: Hampton Press.

Cherry, R. D. & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141), Cresskill, NJ: Hampton Press.

Conlan, G. (1986). "Objective" measures of writing ability. In K. L. Greenberg, H. S Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 109-125), New York: Longman.

Cooper, C. R. & Odell, L. (1977). *Evaluating writing: Describing, measuring, judging*. Urbana, IL: National Council of Teachers of English.

Cremin, T. (2006). Creativity, uncertainty, and discomfort: Teachers as writers. *Cambridge Journal of Education, 36* (3), pp. 415-433.

Crocker, L. (2005). Teaching for the test: How and why test preparation is appropriate. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 159-172). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Cuban, L. (2004). Looking through the rearview mirror at school accountability. In K. A. Sirotnik (Ed.), *Holding accountability accountable: What ought to matter in public* B. *education* (pp. 18-34). New York: Teachers College Press.

Diederich, P. B. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Driscoll, M. P. (2000). *Psychology of learning and instruction* (2nd ed.). Boston: Allyn and Bacon.

Faigley, L., Daly, J. A. & Witte, S. P. (2001). The role of writing apprehension in writing performance and competence. *Journal of Educational Research, 75*(1), 16-21.

Ferrara, S. & DeMauro, G. E. (2006) Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 579-621) Westport, CT: Praeger Publishers.

Firestone, W. A., Monfils, L., Camilli, G., Schorr, R. Y. & Hicks, J. E. (2002). The ambiguity of test preparation: a multi-method analysis in one state. *Teachers College Record, 104*(7) 1485-1523.

Fisher, R., Brooks, G., & Lewis, M. (2002). *Raising standards in literacy*. New York: Routledge.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrra, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*(2), 195-208.

Florida Department of Education (2005). *FCAT Handbook: A Resource for Educators*. Retrieved October 14, 2007 from http://fcat.fldoe.org/handbk/complete.pdf

Florida Department of Education (2006). *Florida Comprehensive Assessment Test 2006 Writing+ Training Materials: Anchor Papers and Qualification Sets* [Computer Software].

Florida Department of Education (2007). *Florida Comprehensive Assessment Test 2007 Writing+ Training Materials: Anchor Papers and Qualification Sets* [Computer Software].

Florida Department of Education (2007). *Keys to FCAT: Information about the 2008 test.* Tallahassee, FL: Florida Department of Education Assessment and School Performance.

Florida Department of Education (2007). *FCAT writing scores* [Data file]. Available from

Florida Department of Education Web site, http://fcat.fldoe.org.

Giordano, G. (2005). *How testing came to dominate American schools: The history of educational assessment*. New York: Peter Lang.

Goodman, D. & Hambleton, R. K. (2005). Some misconceptions about large-scale educational assessments. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 91-110). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Graham, S., Harris, K. R., Fink, B. & MacArthur, C. A. (2001). Teacher efficacy in writing: A construct validation with primary grade teachers. *Scientific Studies of Reading, 5*(2), 177-202.

Gredler, M. E. (2001). *Learning and instruction: Theory into practice* (4th ed.). Columbus, OH: Merrill Prentice Hall.

Heck, R. H. & Crislip, M. (2001). Direct and indirect writing assessments: Examining issues of equity and utility. *Educational Evaluation and Policy Analysis, 23* (3), 275-292.

Higgins, B., Miller, M. & Wegmann, S. (2007). Teaching to the test…not! Balancing best practice and testing requirements in writing. *The Reading Teacher, 60*(4), 310-319.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th Ed). Boston: Allyn and Bacon.

Hoy, A. W. & Spero, R. B. (2005). Changes in teacher efficacy during the early years of teaching: A comparison of four measures. *Teaching and Teacher Education, 21*(6), 343-356.

Huot, B. A. (2002). *(Re) articulating writing assessment for teaching and learning*.

Logan, Utah: Utah State University Press.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206-232), Cresskill, NJ: Hampton Press.

International Reading Association and National Council of Teachers of English Joint Task Force on Assessment (1994). *Standards for the assessment of reading and writing.*

Lane, S. & Stone, C. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 387-431) Westport, CT: Praeger Publishers.

Langford, P. E. (2005). *Vygotsky's developmental and educational psychology*. New York: Psychology Press.

Lederman, M. J. (1986). Why test? In K. L. Greenberg, H. S Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 35-43), New York: Longman.

Linn, R. L. (2000). Assessment and accountability. *Educational Researcher, 29* (2), 4-15.

Lumley, D. R. & Yan, W. (2001, April). *The impact of state mandated, large-scale writing assessment in Pennsylvania.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Lunsford, A. A. (1986). The past – and future – of writing assessment. In K. L. Greenberg, H. S Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 1-11), New York: Longman.

Lyman, H. B. (1998). *Test scores and what they mean* (6th ed.) Boston: Allyn and Bacon.

Mabry, L. (2004). Strange, yet familiar: Assessment-driven education. In K. A. Sirotnik (Ed.), *Holding accountability accountable: What ought to matter in public education* (pp. 116-134). New York: Teachers College Press.

Mabry, L. (1999). Writing to the rubric. *Phi Delta Kappan, 80* (9), 673-679.

Maki, Peggy L. (2004). *Assessing for learning: Building a sustainable commitment across the institution.* Sterling, VA: Stylus Publishing, LLC.

McLaughlin, M. W. & Talbert, J. E. (2006). *Building school-based teacher learning communities: Professional strategies to improve student achievement*. New York: Teachers College Press.

Montgomery, K. (2000). Classroom rubrics: Systemizing what teachers do naturally. *ERIC Clearing House,* 73(6), 324-328.

Moore, A. D., & Young, S. (1997, October). *Clarifying the blurred image: Estimating the inter-rater reliability of performance assessments.* Paper presented at the Annual Meeting of the Northern Rocky Mountain Educational Research Association, Jackson, WY.

Myerberg, N. J. (1996, April). *Inter-rater reliability on various types of assessments scored by school district staff.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

National Writing Project & Nagin, C. (2003). *Because writing matters: Improving student writing in our schools.* San Francisco: Jossey-Bass.

Neill, M. (2000). Old tests in new clothes. *Instructor, 109* (5), 31-34.

Northwest Regional Educational Laboratory (2004). 6 + 1 Trait Writing – About. Retrieved May 15, 2008, from http://www.nwrel.org/assessment/about.php?odelay=1&d=1

Orange County Public Schools (2008). *OCPS District K-12 Writing Plan.* Retrieved May 28, 2008, from https://www.ocps.net/cs/services/Curriculum/languagearts/Pages/WritingProgram.aspx

Ornstein, A. C. & Hunkins, F. P. (1998). *Curriculum: Foundations principles and issues* (3rd ed.). Boston: Allyn and Bacon.

Penny, J, Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7* (2), 143-164.

Phelps, R. P. (2005). Persistently positive: Forty years of public opinion. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 1-22). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Phelps, R. P. (2006). Characteristics of an effective student testing system. *Educational Horizons*, 19-29.

Piaget, J. (1965). The stages of the intellectual development of the child. In B. A. Marlowe & A. S. Canestrari (Eds.), *Educational psychology in context: Readings for future teachers* (pp. 98-106). Thousand Oaks, CA: Sage Publications.

Pilcher, J. K. (2001, March). *The standards and integrating instructional and assessment practices.* Paper presented at the annual meeting of the American association of colleges for teacher education, Dallas, TX.

Pula, J. J. & Huot, B. A. (1993). A model of background influences on holistic raters. In

M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265), Cresskill, NJ: Hampton Press.

Quinlan, A. M. (2006). *A complete guide to rubrics*. Lanham, MD: Rowman & Littlefield Education.

Reckase, M. D. (1997). *Statistical test specifications for performance assessments: Is this an oxymoron?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Reeves, D. B. (2004). *Accountability for learning: How teachers and school leaders can take charge.* Alexandria, VA: Association for Supervision and Curriculum Development.

Rieber, R. W. (ed.). (1997). *The collected works of L. S. Vygotsky: The history of the* development of higher mental functions (Vol. 4). New York: Plenum Press.

Routman, R. (2000). Conversations: Strategies for teaching learning and evaluating. Portsmouth, NH: Heinemann.

Salies, T. G. (1998). *Towards communicative measurement of writing: Where are we now?* U. S. Department of Education Office of Educational Research and Improvement. Retrieved March 24, 2008 from the ERIC database.

Salpeter, J. & Foster, K. (2000). Playing the testing game. *Technology and Learning, 20*, (11), 26-34.

Seminole County Public Schools (2007). *K-12 Comprehensive Writing Plan.* Lake Mary, FL: Seminole County Public Schools.

Shale, D. (1996). Essay reliability: Form and meaning. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 76-96), New

York: The Modern Language Association of America.

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 79-108), Cresskill, NJ: Hampton Press.

Trimbur, J. (1996). Response: Why do we test writing? In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 45-48), New York: The Modern Language Association of America.

Tschannen-Moran, M. & Hoy, A. W. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education, 23*, 944-956.

Van Noord, R. G. & Prevatt, F. F. (2002). Rater agreement on IQ and achievement tests effect on evaluations of learning disabilities. *Journal of school psychology*, *40*(2), 167-176.

Waltman, K., Kahn, A. & Koency, G. (1998). *Alternative approaches to scoring: The effects of using different scoring method on the validity of scores from a performance assessment.* U. S. Department of Education Office of Educational Research and Improvement. Retrieved March 24, 2008 from the ERIC database.

Weigle, S. C. (2002). *Assessing Writing*. New York: Cambridge University Press.

White, E. M. (1986). Pitfalls in the testing of writing. In K. L. Greenberg, H. S Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 53-78), New York: Longman.

White, E. M. (1993). Holistic scoring: Past triumphs, future challenges. In M. M.

Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 79-108), Cresskill, NJ: Hampton Press.

Wiggins, G. (2006). Teaching to the (authentic) test. In B. A. Marlowe & A. S. Canestrari (Eds.). *Educational psychology in context: Readings for future teachers* (pp. 252-263). Thousand Oaks, CA: Sage Publications.

Willett, T. (2001). *English holistic assessment validation*. U. S. Department of Education Office of Educational Research and Improvement. Retrieved November 20, 2007 from the ERIC database.

Williamson, M. M. (1993). An introduction to holistic scoring: The social, historical, and theoretical context of writing assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 1-43), Cresskill, NJ: Hampton Press.

Winkler, A. (2002). Division in the ranks: Standardized testing draws lines between new and veteran teachers. *Phi Delta Kappan, 84,* (3), 219-225.

Wint-Tat Chiu, C. (2001). *Scoring performance assessments based on judgments: Generalizability theory.* Boston: Kluwer Academic Publishers.

Yancey, K. C. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication,* 50 (3), 483-503.

Zimmerman, B. J. & Schunk, D. H. (2003). Albert Bandura: The scholar and his contributions to educational psychology. In B. J. Zimmerman & D. H. Schunk (Eds.), *Educational psychology: A century of contributions* (pp. 431-457), Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.