ACCOUNTABILITY AND TEACHER ATTITUDES: CONSEQUENTIAL VALIDITY EVIDENCE FOR FLORIDA'S EDUCATIONAL ACCOUNTABILITY SYSTEM

By

KATHRYN ELIZABETH MILLER

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS IN EDUCATON
UNIVERSITY OF FLORIDA

2006

To the LaFrance women

ACKNOWLEDGMENTS

I would like to thank my committee members (Dr. M. David Miller and Dr. Anne Seraphine) for guiding me through each semester, fielding all my questions throughout my graduate career. I would also like to thank my fellow students (Jann Macinnes, Jenny Bergeron, and Janna Underhill) for all the study groups they pioneered. If not for their complete dedication to academics, my experience would have been greatly diminished. I would also like to thank Elaine Green and Linda Parsons for keeping me sane during my teaching assistantships.

Many other people have offered support or motivation, or simply inspired me in innumerable ways. Special thanks go to Joshua Marland, Crystal Calkins, Janna Baumann, Samuel Hanna, Sally King, Mico Adorno, Laura McCoy, Amy Godfrey, Allison Knowlton, Andrew Brunelle, Philip Moring, and Anthony Herman. I would also like to thank my mother, Jacquie Hernandez, for her unconditional love and support. I thank my brother, Steven Miller, and my sister, Melissa Miller. They make life fantastic and challenge me at every turn. My Aunt, Gisele Andrade, has always been a role model of mine. Her strength is the motivating force in my life.

TABLE OF CONTENTS

page

LIST OF TABLES

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Arts in Education

ACCOUNTABILITY AND TEACHER ATTITUDES: CONSEQUENTIAL VALIDITY EVIDENCE FOR FLORIDA'S EDUCATIONAL ACCOUNTABILITY SYSTEM

By

Kathryn Elizabeth Miller

May 2006

Chair: David Miller
Major Department: Educational Psychology

Consequences that arise from the Florida Comprehensive Assessment Test (FCAT), in accordance with the No Child Left Behind Act of 2001 (NCLB), impact the overall validity of the accountability design established by Florida. In developing a validity argument for uses and interpretations of the FCAT, 75 teachers employed by six different schools responded to a survey ascertaining their opinions on NCLB, Florida's A+ Plan, the Sunshine State Standards, and the FCAT. Data were analyzed looking at the areas of accountability as separate entities and the attitudes towards each area. Data also were analyzed to examine differences across teachers, to uncover factors that may influence a teachers' view of accountability.

CHAPTER 1
INTRODUCTION

The No Child Left Behind Act of 2001 (NCLB) marked the beginning of a new era in education, where each state became responsible for creating a system of educational accountability. Accountability refers to the process of holding school districts, schools, teachers, and students responsible for learning. An accountability system is a structure for making decisions and applying consequences based on information collected from assessments. Decisions and consequences that result from accountability systems range from allocation of funds, to third-grade promotion. The high-stakes nature of these accountability designs necessitates a thorough examination of the validity and the assessments they encompass. Our purpose was to gather information that can be used toward an argument for validity.

Validity is an overall appraisal of the degree to which an assessment's use and interpretation are adequate and appropriate (Messick, 1995). A validity argument is made by collecting empirical evidence and providing theoretical rationales for the uses and interpretations (Haertel, 1999).

We focused on providing evidence for a validity argument for Florida's use of the Florida Comprehensive Assessment Test (FCAT). Developing a validity argument is multifaceted and includes examining consequences that arise from high-stakes testing. Our study mainly addressed the new accountability legislation and its consequences for teachers in the state of Florida.

To better illustrate Florida's accountability system, factors that contribute to it or are included in it are outlined in the next sections. NCLB is of interest because it is the new legislation that must be adhered to by each state. In fulfilling NCLB, Florida's system includes measuring Adequate Yearly Progress (AYP), the A+ Plan, the Sunshine State Standards (SSS), and the FCAT.

National and State Accountability

No Child Left Behind Act of 2001

NCLB is at the center of educational accountability in every state. NCLB is the new version of the Elementary and Secondary Education Act written into law in 1965 and provides billions of dollars in federal funding for various educational programs (USDOE, 2006a). The purpose of NCLB is to ensure that every child in America is able to meet the high learning standards of the state the child resides in. The act, though intricate and complex, is founded on four basic principles; stronger accountability, increased flexibility and local control, more options for parents, and emphasis on proven teaching methods (USDOE, 2006b). NCLB aims at improving education all over the United States and raising the bar for what is deemed acceptable learning. Goals of NCLB are numerous, specific, and lofty. The goals most pertinent (No Child Left Behind, 2002) to our study are listed below.

- All students will reach high standards, at a minimum attaining proficiency or better in reading and mathematics by 2013-2014.
- By 2013-2014, all the students will be proficient in reading by the end of the third grade.
- All limited English proficiency (LEP) students will become proficient in English.
- All students will be taught by highly qualified teachers.

- All students will be educated in learning environments that are safe, drug-free and conducive to learning.
- All students will graduate from high school.

NCLB requires that each state develop its own accountability system that is valid, reliable, and meets all requirements outlined in the act. The degree to which each system is valid and reliable is individually established by each state. The 2005-2006 school year marked the deadline for testing all students grades 3-8 in mathematics and reading, annually. Science must be included in the testing regime by the 2007-2008 school year, at least once during elementary, middle, and high school. All assessments must be aligned with the content standards established by the state. All students must be proficient by 2013-2014. Each state determines its own guidelines for proficient status (Lane, 2004). The 2002-2003 school year marked the deadline for each state to furnish annual report cards of their progress. The report cards include information on student achievement by district and subgroup. Minority students, students with disabilities, LEP students, and children from low-income families are all included in the annual report cards (No Child Left Behind, 2002).

Florida's accountability system, in fulfilling NCLB, includes AYP, Florida's A+ Plan (school grades), individual student progress towards (or consistent proficient levels of) mastery on the FCAT, and a return on investment. Return on investment is a measure that relates dollars spent to student achievement (FDOE, 2005b). These elements are designed to provide a cohesive and extensive representation of a school's performance and are made available for parents, educators, and members of the community.

Adequate Yearly Progress as Determined in Florida

"Adequate Yearly Progress measures the progress of all public schools, and school districts toward enabling all students to meet the state's academic achievement standards" (FDOE, 2005b, p.1). AYP targets the performance of every subgroup and aims to ensure that in 1 year's time, students are learning 1 year's worth of knowledge as delineated in the content standards. Subgroups are created on the basis of race or ethnicity, socioeconomic status (SES), disability, and English proficiency. States are required to define AYP for the state, school districts, and schools in a way that will facilitate all students to meet the state's achievement standards by 2014 (FDOE, 2005b).

Florida uses the FCAT to ascertain each student's level of proficiency, a necessity for making AYP. There are five possible achievement levels one can attain from the single FCAT score. The levels range from 1 to 5. Level 1 is below basic, Level 2 is basic, Levels 3 and 4 are proficient, and Level 5 is advanced. All students scoring a 3 or above are considered proficient for classification purposes. Florida also has a separate assessment for students with disabilities who would not be able to earn a standard diploma. The Florida Alternate Assessment Report (FAAR) also uses a 5-point scale to determine a student's proficiency level of the SSS. The FAAR scale is as follows; Levels 0 or 1 are below basic, Level 2 is basic, Level 3 is proficient, and Level 4 is advanced (FDOE, 2005b).

For a school in Florida to make AYP, 95% of all students and all identified subgroups must partake in the FCAT or alternative assessment (when applicable). A subgroup must include at least 30 students to be included in the AYP calculations. All goals must be met by 2014 and a blue print for the progression toward that goal must also be agreed on and met annually (FDOE, 2003). For example, if Florida declares that 68%

of students will be proficient in mathematics by 2007, then that goal must be realized to make AYP. Additionally, there must be a 1% increase in the percentage of students proficient in writing. If the annual objectives for reading or mathematics are not met by subgroups in a school or district, AYP can still be met if the percentage of nonproficient students decreased by 10% from the previous year. It is not possible to make AYP if (under Florida's A+ Plan) a school receives a D or F (FDOE, 2003).

Florida's A+ Plan

Florida's A+ Plan is a grading system for schools: A is the highest grade a school can receive and F is the lowest (the A+ plan uses a traditional grading scale of A, B, C, D, and F). To make AYP, a school must first receive a grade of C or higher (FDOE, 2005d).

- A = 410 points or more, meet AYP of bottom 25% in reading, gains for bottom 25% are within 10 points of gains for all students, and 95% of eligible students are tested
- B = 380 points or more, meet AYP of bottom 25% in reading within 2 years, and 90% of eligible students are tested
- C = 320 points or more, meet AYP of bottom 25% in reading within 2 years, and 90% of eligible students are tested
- D = 280 points or more and 90% of eligible students are tested
- F = Fewer than 280 points or less than 90% of eligible students tested.

A school can earn points if their students do well on the assessments or improve from the previous year. Schools earn one point for every percent of students scoring 3, 4, or 5 in mathematics. Schools also earn will also receive one point for every percent of students scoring a 3, 4, or 5 in reading. One point is also given for each percent of students who score a 3 or above on the writing assessment. For each percent of students who gain one achievement level and for students who maintain a level of 3 or above, one

point is awarded. One point is awarded for each percent of students in Levels 1 or 2 demonstrating more than one year's growth. One point is awarded for each percent of the lowest performing readers (bottom 25%) making learning gains from the previous year (FDOE, 2005d).

Sunshine State Standards

The SSS were approved by the Board of Education in Florida in 1996. The standards provide expectations for student performance and achievement. The standards were written in seven subject areas and aimed to allow flexibility in curriculum, catering to the needs of different schools. In recent years, changes to the SSS were made to better accommodate new accountability legislation. Grade level expectations for the major subject areas were added and are guidelines for the FCAT. The subject areas outlined in the SSS are music and fine arts, foreign language, language arts, mathematics, science, social studies, physical education, and health (FDOE, 2005e).

Florida Comprehensive Assessment Test

The FCAT is the only test administered statewide that is designed to align with the SSS. The FCAT determines a students' level of achievement at each grade level making it the primary AYP determinate. The FCAT has two major components, a norm-referenced test (NRT) and a criterion-referenced test (CRT). The NRT currently being used as part of the FCAT is the Stanford 10. It is used to compare individual students in Florida to national norms (FDOE, 2005c). NRT's are designed to maximize response variance and artificially spread the scores (Kohn, 2000). Reading and mathematics are the only subjects measured in the NRT.

The CRT is designed to measure a students' level of mastery of the SSS in reading, writing, mathematics, and science. These scores are not designed to be used for

comparison, but as a comprehensive exam measuring knowledge gained inside the classroom. All students in grades 3-10 are required to take the FCAT Reading and Mathematics. Students in grades 4, 8, and 10 also take the FCAT Writing and grades 5, 8, and 11 take the Science portion of the FCAT (FDOE, 2005c).

The FCAT is a high-stakes assessment because of the consequences attached to the scores (Kohn, 2000). The results impact grade to grade promotion, funds allocated to schools, high school graduation, teacher rewards, and how the school is viewed by the community. Intuitively, the stakes attached to the FCAT put many teachers under an extreme amount of pressure to increase student performance. The pressures felt by teachers and their opinions have not been extensively studied. Our study aims to examine the consequences of high-stakes testing, for teachers.

Effects of Testing on Teachers

The fundamental rationale for our study is to give part in the validation of Florida's accountability system as a whole by examining teacher opinions. Our study examines opinions for all areas in the accountability design, focusing primarily on the uses and interpretations of the FCAT. Validation is essential because of the consequences attached to FCAT scores. AYP and school grades are directly determined by FCAT scores. Schools that do not make AYP have serious consequences. To illustrate, after 5 years of failing to make AYP a school will be identified for restructuring. Restructuring entails implementing "significant alternative governance actions, state takeover, the hiring of a private management contractor, converting to a charter school, or significant staff restructuring" (FDOE, 2005b). These and other consequences are the driving force behind our study. Teacher opinions about the validity of each area of accountability will be surveyed, specifically FCAT content, SSS, Florida's A+ Plan, and NCLB.

CHAPTER 2
LITERATURE REVIEW

The significance of a methodical examination of validity is emphasized in measurement literature and widely studied by assessment specialists. Proper validation methodology and threats to validity are outlined throughout this literature review providing a foundation for our study.

Validity Argument

The concept of forming a validity argument is that "validation should be a process of constructing and evaluating arguments for and against proposed test interpretations and uses" (Haertel, 1999, p. 5). In a meeting of the National Council on Measurement in Education (NCME) in 1999, President Edward Haertel explained common flaws in the validation of an assessment's use. Planning a validity argument is often done by going down a checklist. Checking off items shows what has been accomplished and leaves little room for discovering evidence against the intended interpretation (Haertel, 1999). According to Cronbach, "the task of validation is not to uphold a test, practice, or theory. Ideally, validators will prepare as debaters do. Studying a topic from all angles, a debater grasps the arguments pro and con so well that he or she could speak for either side" (Cronbach, 1988, p.3). Haertel points out that though validation, in practice, may be flawed, few people are willing to investigate or change the uses and interpretations of tests (Haertel, 1999).

Validity is the degree of appropriateness and adequacy for the intended use and interpretation of an assessment. According to Messick (1995), "Validity is not a property

of the test or assessment as such, but rather the meaning of the test scores". Validity is a unitary concept and evidence must be collected from different perspectives. There are six aspects of validity to consider when developing a validity argument emphasizing content, substantive, structural, generalizability, external, and consequential basis for construct validity (Messick, 1995). A brief overview of ways to collect evidence for each aspect of validity is followed by major validity issues surrounding high-stakes testing and validation practices in general.

- Collecting evidence for the content aspect of validity includes content relevance, proper representation of learning objectives to be measured and items, and over item quality.
- Collecting evidence for the substantive aspect of validity refers to test respondents engaging in the proper mental processes required by each assessment task.
- Collective evidence for the structural aspect of validity includes the extent to which the internal structure of the assessment, individual items, and scoring rubrics align with construct domain of interest.
- Collecting evidence for the generalizability aspect of validity studies the extent to which score interpretations generalize to and across different population groups, settings, and tasks.
- Collecting evidence for the external aspect of validity is the process of using already established tools and practices to judge the quality of the new assessment or system.
- Collecting evidence for the consequential aspect of validity consists of discovering the consequences, both actual and potential.

We were concerned primarily with consequential evidence, namely the effects of high-stakes testing on teachers. After teachers receive the results, they can lend judgment to the comparability of the test scores and the students abilities during class time. Also, teachers see the direct effects of testing on students before, during, and after testing times, hence affording another viewpoint on the authority of the FCAT's use. The scores generated from the FCAT are comprised of the knowledge acquired in class and many

other factors. The FCAT CRT measures specific domains of knowledge acquired in class, factors contributing to the students score external to each domain are forms of error (Haladyna & Downing, 2004). Potential contributing factors to error are discussed in the next sections. These are important considerations for our study because teachers are observers and contributing sources, of error.

Validity Issues in High-Stakes Testing

When high stakes are attached to an assessment, test developers, school officials, and decision makers ensure several aspects of validity. The reliability indices will undoubtedly be above .9 or .95. The content measured in the assessments will be directly drawn from the standards set forth by the state and taught in every classroom. Despite the attention placed on validation, numerous problems arise in high-stakes testing. The next sections describe some validity issues examined in our study.

Construct-Irrelevant Variance

Construct-irrelevant variance (CIV) is systematic error variance or bias. An examination of the contributing sources of CIV is important for our study because teachers are instructors, test preparers, and test administrators. Teachers have the propensity to impact CIV in a myriad of ways, thereby shaping validity.

Lord and Novick (1968, p.43) describe “systematic error as an undesirable change in true score”. Systematic error correlates to both true and observed scores because each individual within the group is either affected on unaffected by the CIV (Haladyna & Downing, 2004). To illustrate, a student scores a 130 on an IQ test that has a standard error of measurement (SEM) of 3. According to classical test theory, the SEM is derived mathematically, consistent across test takers, and accounts for random error (Crocker & Algina, 1986). Other factors may have been measured systematically into their score

having nothing to do with the construct of intelligence. Anything besides the construct of interest possibly measured with the construct is CIV (Haladyna & Downing, 2004).

Contributing sources of CIV can be specific to an individual or group. An example of systematic error that is constant for an entire group is a rater that is more strict than their colleagues. If the rater administering the IQ test scores too stringently, they are contributing to systematic error, measured with the true score of the student. Every student assessed by that rater will be at a slight disadvantage. Also, when there are multiple forms of a test, it is always possible that one form will be slightly more difficult than the rest. The entire group receiving the more difficult form will have a score that is an underestimate of their true score for that particular construct. Likewise, an entire group may have an easier form and their results will be an overestimate of their true ability (Haladyna & Downing, 2004).

The other type of error that occurs systematically is specific to individuals. Perhaps the most common source of CIV is reading comprehension (Haladyna & Downing, 2004). This occurs when the students' ability to read the question affects their answer. For example, a student may know the answer to a question about the solar system, but because of the vocabulary in the question, are unable to answer. This is especially problematic for LEP students (Abedi, 2004). If the student would score higher on an identical form of the test written in their native language, CIV is affecting their results.

Understanding CIV is important for our study because of the implications for FCAT results. When interpreting FCAT scores, it is important to consider all potential elements measured in the raw score. Every student has different innate abilities,

motivations, and distractions measured in their results. It is important to acknowledge and minimize error, thereby increasing validity.

Test Preparation

Test preparation is recommended by assessment specialists. Preparation influences the error variance in test results. Sound preparation includes providing examples of different item formats, motivating students, teaching students to use time effectively, making educated guesses, and so on. Students that are properly prepared for a test will do better than students without preparation. However, it is possible to prepare students too much for an exam. The only way to prevent CIV is for each district, school, and educator to uniformly prepare their students according to the guidelines provided to them in the testing manuals (Haladyna and Downing, 2004).

Beyond uniformly preparing students, Haladyna and Downing (2004) discuss the ethical issues which arise from high-stakes test preparation. They address specific issues including curriculum developed on the basis of test content as opposed to content standards established by the state, providing students with similar or identical items, or anything that may narrow the intended curriculum. High stakes tests like the FCAT are designed to draw a representative sample from a larger domain and assess it. Students should be taught all of the domain (or content standards) and not overly exposed to information that is more likely to be on the FCAT.

If the construct is an ability (rather than a domain of knowledge), different problems may occur. The FCAT Writing is a writing assessment administered each year to students in grades 4, 8, and 10. If students are taught to write in accordance to the FCAT Writing rubric and are not exposed to other styles of writing, it would be an

example of construct-irrelevant easiness (Haladyna 2004). The score from the writing assessment will give an inflated estimate of the students writing ability.

Sources of Unreliability for the No Child left Behind Accountability Designs

In the previous sections, possible threats to the validity of specific assessments, high-stakes testing in general, and accountability designs as a whole were discussed. For a viable degree of validity to exist, some reliability (or consistency) must be present. Reliability is most commonly examined as a property of an assessment and not for an entire accountability design (Hill & DePascale, 2003).

The NCLB act requires each subgroup within a school to make AYP. Twelve states have established a cut-off group size that they deem reliable. The cut-offs for those 12 states range from 10 to 75 students, with a median of 30. Florida requires a minimum of 30 students in a subgroup to be counted (FDOE, 2003). The cut-offs are in place to ensure results collected yield reliable information about a subgroup. For example, if three Native-American students attend one school, it is impossible to get any reliable information from their test results. A general reliability rule is the more information (test results) the higher the reliability.

NCLB requires all subgroups make AYP. One subgroup can cause an entire school to fail, reinforcing the need to ensure the reliability (and validity) of the accountability design. The recommended number of students required for each subgroup is much higher, than in practice. Hill and Depascale (2003) suggest that roughly 300 students would be adequate. This would encompass very few subgroups, greatly diminishing the validity of the accountability design. To reiterate, for the results to be reliable, the number of students needed would be far larger than most subgroups. Only testing larger subgroups

diminishes the validity of the accountability design and negates the entire purpose of NCLB.

Positive Consequences of High-Stakes Testing

The general sentiment towards high-stakes testing in measurement literature is unenthusiastic, but there are positive effects of testing. Our study examined intended, unintended, positive, and negative consequences of testing and their effects on teachers. Positive consequences of accountability examined in our study are outlined below. Cizek (2001) described the following 10 consequences in an article about unintended consequences.

- Professional Development- Professional development for educators has been "spotty, hit or miss, of questionable research base, of dubious effectiveness, and thoroughly avoidable" in the past and sometimes at present. However, professional development is becoming increasingly better over time. The new accountability policies and "Principles of High-Quality Professional Development" established by the Department of Education are ensuring teachers are constantly gaining new knowledge and expertise in their subject areas.

- Accommodation- The new federal legislation requires that all students be tested. All students must be assessed and accommodated. Extra attention is given to students who need it and much focus has been brought to students who may have been overlooked in the past. Cizek mentions a research study where disadvantaged students, who had some history of failing, reported that their teachers began to focus more attention on them after the high-stakes testing and accountability program was established.

- Knowledge About Testing- The constant submersion in high-stakes testing has aided in educating teachers on test content, consequences, and construction. Teachers understand the entire practice of testing more now than ever. This can affect how well they write tests, grade exams, develop rubrics, and their assessment practices in general at the classroom level.

- Collection of Information- School districts have become more conscientious about their data collection practices.

- Use of Information- The accountability movement is in full swing which means finding information about test scores, funding, spending, graduation rates, and the like, is as easy to pull up over the internet as your favorite recipe. This information is all used to improve programs and allocate funds where needed.

- Educational Options-In addition to traditional public schools, parents and students often have the option of charter schools, magnet schools, and home schooling.

- Accountability Systems-Cizek argues that high-stakes tests are often the foundation for accountability systems and that accountability in its connotation today is because of high-stakes testing.

- Educators Intimacy with their Discipline-The idea behind this consequence is that educators chosen to be involved with content or test development will be immersed in discussion about the content and it will trickle down to the local level.

- Quality of Tests-Tests today are "highly reliable, free from bias, relevant and age appropriate, higher order, tightly related to important and public goals, time and cost efficient, and yielding remarkably consistent decisions" according to Cizek (2001).

- Increased Student Learning- The primary goal and intended consequence of high-stakes testing is to increase student learning. There is research that shows a positive relationship between the presence of high stakes testing and student scores on the International Assessment of Educational Progress in Canada. In addition there are other studies that show favorable results for high-stakes testing.

Collecting consequential evidence for validity is the primary focus of our study. Suggestions made by Cronbach (1988), Messick (1995), and Haertel (1999) for forming a validity argument will be ensued throughout our study. Also, the survey instrument was developed and analyzed based on the information and validity cautions provided in the articles by Haladyna & Downing (2004), Abedi (2004), Hill & Depascale (2003), and Cizek (2001).

CHAPTER 3
METHODS

Respondents

The sample of 75 teachers were returned from the experimentally accessible population of 261 teachers employed in six elementary schools (School A: n = 30, School B: n = 50, School C: n = 55, School D: n = 33, School E: n = 52, School F: n = 41) from two school districts, one in central Florida and five in north-central Florida. The original protocol was to investigate six elementary schools from the same district. These schools were selected based on their accountability success (school grade in the A+ plan) and AYP status from 2004. Permission to survey teachers was sought from schools receiving grades of A, B, C, or D. There were no F schools in this school district. A representative sample of schools was sought to compare the views of teachers from schools of varying success with Florida's accountability system.

The theoretical premise behind choosing schools receiving both high and low grades was to gain a lucid depiction of the consequences of testing at the classroom level and the opinions of teachers from dissimilar schools in relation to each other. For example, schools with a poor accountability record (low grades), may place more stress on teachers to improve their students' FCAT scores. Also, teachers from schools having no success with accountability may be more apathetic than teachers from schools with established success. Likewise, teachers from successful schools may be under constant pressure to improve or maintain high FCAT scores.

Five out of the six schools approved the protocol. The school that declined was replaced by a school with a similar accountability record (School C). However, the school is located in a different school district. Table 3-1 displays demographic and accountability information for each school (FDOE, 2005a). Seventy-five (28.7%) surveys were returned within the allotted timeframe [School A: *n* = 15 (50%), School B: *n* = 9 (18%), School C: *n* = 10 (18.2%), School D: *n* = 5 (15.2%), School E: *n* = 22 (42.4%), School F: *n* = 14 (34.2%)].

All teachers from kindergarten through fifth grade were asked to participate including Exceptional Student Education (ESE), Gifted, English for Speakers of Other Languages (ESOL), Physical Education (P.E.), Art, Music, and Speech. Administration was not asked to participate.

Table 3-1 School Demographic Information

School	Location	Grade 2005	Grade 2004	Grade 2003	Total Students	SES %	Minority %	AYP Status
A	N. central	A	A	B	219	54	31	Provisional
B	N. central	D	D	C	183	93	97	Not met
C	Central	A	A	A	396	39	37	Met
D	N. central	D			88	92	95	Not met
E	N. central	B	B	A	329	41	39	Provisional
F	N. central	D	C	B	177	87	84	Not met

*SES is based on percentage of students eligible for free and reduced lunch

Materials

A survey instrument was developed to determine teacher opinions of accountability at the national and state levels, mainly the effects of the new laws on themselves and their students. In addition to measuring opinions held by teachers, this survey was developed to address certain validity concerns that are influenced by teachers in terms of gathering

consequential evidence for a validity argument, such as teaching to the test (i.e., teachers will be asked to what extent they stress material that is likely to show up on the FCAT).

The survey has three parts and consists of three questions and 34 statements with a corresponding 5-point Likert scale. The scale is from “strongly disagree” to “strongly agree” and contains a neutral point. The first section comprises statements (items 1-11) about the NCLB act, Florida’s A+ Plan, and general items about Florida’s accountability design. Part two ascertains opinions (items 12-34) pertaining to Florida accountability design on a less macro level, particularly the SSS and the FCAT. Teachers have a more intimate relationship with the SSS and the FCAT, so more items (and of greater detail), were included in this section. The third section contains three open-ended questions inquiring professional information about the participant (Appendix shows entire survey).

Procedure

Once permission was given by each school, the surveys were hand-delivered, along with an invitation to participate and a self-addressed stamped envelope for each teacher. Surveys were color-coded by school for identification purposes. Packets containing the above mentioned items were placed in teacher mailboxes by school personnel for teachers to examine at their leisure. Teachers were given written instructions to return the surveys within a specified timeframe, approximately 2 weeks, on average for each school. The length of time it took to gain permission from schools varied extensively causing the packets to be delivered on different days between the months of August and October in 2005.

Analysis Approach

The design was based on establishing five independent variables, or five separate areas of accountability that are in the realm of a teachers’ expertise. The five branches of

interest are (corresponding new variable labels are in parenthesis): (1) No Child Left Behind Act of 2001 (SUMNCLB) (2) Florida's A+ Plan (SUMAPLAN) (3) the Sunshine State Standards (SUMSSS) (4) the Florida Comprehensive Assessment Test (SUMFCAT) (5) the subsections of the FCAT (SUMFCATSECT). Each variable was created by summing the responses of like items on the survey instrument. Grouping different items to formulate new variables secures a more reliable measure of an overall attitude towards a specific subject. In addition, a summated score for each participant was calculated and used in the analysis. The summated score (AVERAGE) was used as a comprehensive measure for each individual's stance on accountability and derived from items that specifically addressed an attitude. Cronbach's alpha was computed for each new variable as a measure of reliability (Table 3-2).

Table 3-2 Reliability Statistics for New Variables

New Variable	Statements*	N	Cronbach's alpha
SUMNCLB	2, 3, 4	72	.903
SUMAPLAN	5, 6, 8	68	.729
SUMSSS	12, 13, 14	70	.743
SUMFCAT	15, 16, 21	69	.826
SUMFCATSECT	31-34	60	.912
AVERAGE	2-6, 8, 12-16, 21	63	.848

*Statements found in Appendix

The primary function of the survey and purpose of our study was to uncover teacher opinions at definite levels of accountability. Teachers were surveyed in hopes of them lending a unique perspective on the validity of Florida's educational accountability design and the consequences of high-stakes testing. A secondary focus of our study was to examine differences within the sample and uncover factors contributing to the beliefs held by each teacher. In theory, teachers from different schools (i.e., instructing diverse subpopulations of students) should have very different experiences with the practices

measured in the survey. The aim of looking at teachers as subpopulations was to uncover variations in opinions that can be directly influenced by the working/teaching environment. The rationale for subdividing teachers by school was to gain an understanding of how the consequences of high stakes testing affect teachers, from dissimilar schools, in varying respects.

Some variables used in our study occur naturally as a function of sampling or the demographic information provided by the teachers. The variables of interest SCHOOL (school where respondent teaches), GRADE (school grade in the A+ Plan), YEAR (the number of years the respondent has been teaching), INSIGHT (item 6: the grade attached to each school gives parents insight into how well that school is operating), and IMPACT (item 9: I have an impact on the grade my school receives). SCHOOL and GRADE were analyzed as categorical variables on a nominal scale. YEAR, IMPACT, and INSIGHT are quantitative variables, on an interval scale.

Research Question 1: Teachers' opinions of accountability will be significantly different at each of the four areas of accountability: (1) NCLB (2) A+ Plan (3) SSS (4) FCAT

Using the variables created from the existing data set, analyses were performed to check the overall attitudes towards each accountability branch by the entire sample of teachers. SPSS was used to run six non-directional pair-wise dependent samples *t*-test to test this hypothesis. A Bonferroni adjustment will be made to control for family-wise Type 1 error rate (alpha = .05/6). Rejection of the null hypothesis for an individual *t*-test indicates that there is a statistically significant difference between the two areas of accountability.

Research Question 2: GRADE will be a contributing factor to the responses on IMPACT, INSIGHT SUMNCLB, SUMAPLAN, SUMSSS, and SUMFCAT

Planned complex contrasts were performed to check for mean differences, where schools with a grade of A and B will be contrasted to schools with a grade of D on each of the six variables of interest. IMPACT and INSIGHT were selected based on the empirical know-how that teachers from underachieving schools would have different opinions on items that specifically address their school grade (i.e., teachers with low scoring students are less likely to attribute their students' and schools failures to themselves). Mean differences are of interest for SUMNCLB, SUMAPLAN, SUMSSS, and SUMFCAT because it builds on the first research question by breaking down opinions of each area of accountability across teachers by grade. A Bonferroni adjustment was made to control for the family-wise Type 1 error rate.

Research Question 3: YEAR will have a linear relationship with AVERAGE

A simple linear regression was conducted to test whether the two variables have a linear relationship. If the simple model is accepted YEAR can be used, in part, as a predictor for overall teacher attitudes.

Research Question 4: Teachers will rate the subsections of the FCAT statistically higher than they rate the FCAT as a whole

A directional pair-wise dependent samples *t*-test will be performed to test this research question. Rejection of the null hypothesis will indicate that teachers rate the subsections (i.e., mathematics, science, reading, and writing) higher than the FCAT in its entirety, in terms of being an adequate measure of a student's level of mastery. Type 1 error rate will be set at alpha = .05. This test is being conducted to verify teacher attitudes towards the FCAT. In theory, teachers could rate the FCAT as an indicator of a students' level of mastery lower than if it was broken down into subsections.

CHAPTER 4
RESULTS

The descriptive statistics of the measures (i.e., statements on the Likert scale) included in the overall sample are shown in Table 4-1. The mean score is the average of all responses for a particular item in terms of the scale of the item response. For example, because the responses are on a 5-point Likert scale, an average response of 1.58 suggests that on average the responses fell somewhere between "1-strongly disagree" and "2-disagree". A mean of 3.1 indicate the response fell slightly above "3-neutral". It is widely accepted and often recommended (i.e., if the Likert scale has at least five points it can be considered continuous) to analyze this data as if it were interval, technically it is ordinal data.

Table 4-1 Descriptive Statistics for Survey Statements

Statements	N	Mean	SD
In general, the Florida accountability system works well**.**	73	2.73	1.00
Goals set forth by the NCLB Act will most likely be actualized.	73	2.21	1.01
The NCLB Act has an overall positive impact on the United States.	73	2.62	1.05
The NCLB Act has an overall positive impact on Florida.	74	2.57	1.07
The A+ Plan holds schools accountable for their students learning.	70	3.09	1.10
The grade attached to each school helps give parents insight on how well that school is operating**.**	74	2.15	1.12

Table 4-1 (continued)

Statements	N	Mean	SD
Students from low performing schools should be able to transfer to another school.	74	2.97	1.09
The A+ Plan helps motivate teachers and administrators.	71	2.41	1.33
I have an impact on the grade my school receives.	74	3.87	0.91
Administrators have an impact on the grade their school receives.	73	3.86	0.89
The student body has an impact on the grade their school receives	73	4.22	0.95
The SSS adequately outlines the curriculum content at each grade level.	72	3.76	0.86
The SSS lay the foundation for a broad curriculum.	72	3.67	0.96
All the SSS will be taught at one point or another.	70	3.63	0.97
The FCAT measures the SSS well.	69	2.75	0.96
The FCAT assesses the most important material at each grade level.	69	2.59	1.02
Florida is enacting the NCLB Act appropriately with the FCAT.	70	2.47	1.02
The high stakes attached to the FCAT are necessary.	74	1.99	1.17
The FCAT would still be taken seriously if there weren't consequences for students.	73	3.03	1.01
The FCAT would still be taken seriously if there weren't rewards for teachers.	72	3.64	1.03
The FCAT is a good indicator of the student's level of mastery for required curriculum.	72	2.68	0.96

Table 4-1 (continued)

Statements	N	Mean	SD
I have control over how my students perform on the FCAT.	68	2.87	0.99
I spend extra time in class stressing material that is likely to show up on the FCAT.	67	3.96	0.99
I spend more time going over test taking skills now than before the FCAT was established.	65	3.83	1.11
The FCAT has an overall positive impact on Florida.	74	2.35	1.07
The FCAT has an overall positive impact on my school.	73	2.44	1.12
The FCAT has an overall positive impact on my students.	70	2.24	1.12
Item types (e.g. multiple choice) used on the FCAT, are the most appropriate type for each learning objective.	69	2.84	1.07
Prompts used in the writing section of the FCAT are adequate for measuring the student's overall writing ability.	67	2.76	1.07
Other subjects (e.g. social studies, art) should be included in the FCAT.	70	2.36	1.35
The FCAT measures the most important concepts in Reading.	64	3.56	.0.94
The FCAT measures the most important concepts in Writing.	63	3.22	0.99
The FCAT measures the most important concepts in Mathematics.	64	3.42	0.94
The FCAT measures the most important concepts in Science.	60	3.05	0.96

A Likert scale gives the option of using a neutral point. The neutral point allows respondents that are apathetic towards the topic, an opportunity to answer fairly. Unfortunately, using a neutral point can cause the survey results to sway towards the middle and the results often appear to be insignificant. Part of our study intended to uncover indifferent attitudes towards certain aspects of educational accountability. For this purpose a neutral point was used and the percent agreement and disagreement for each statement are shown in Table 4-2. The percentage reported is the valid percent, which does not take into account missing data (i.e., the percent is out of the people who responded to that particular item)

Table 4-2 Percent Agreement Statistics for Survey Statements

Statement	% Agree	% Disagree
In general, the Florida accountability system works well**.**	17.8	39.8
Goals set forth by the NCLB Act will most likely be actualized.	9.5	64.4
The NCLB Act has an overall positive impact on the United States.	17.8	52.1
The NCLB Act has an overall positive impact on Florida.	17.6	54.0
The A+ Plan holds schools accountable for their students learning.	32.8	27.2
The grade attached to each school helps give parents insight on how well that school is operating**.**	12.2	67.6
Students from low performing schools should be able to transfer to another school.	29.8	25.7
The A+ Plan helps motivate teachers and administrators.	24.0	57.7

Table 4-2 (continued)

Statement	% Agree	% Disagree
I have an impact on the grade my school receives.	67.5	8.1
Administrators have an impact on the grade their school receives.	72.6	9.6
The student body has an impact on the grade their school receives	79.4	5.5
The SSS adequately outlines the curriculum content at each grade level.	66.7	7.0
The SSS lay the foundation for a broad curriculum.	62.5	11.1
All the SSS will be taught at one point or another.	57.2	10.0
The FCAT measures the SSS well.	21.7	44.9
The FCAT assesses the most important material at each grade level.	17.3	49.2
Florida is enacting the NCLB Act appropriately with the FCAT.	17.1	52.9
The high stakes attached to the FCAT are necessary.	12.2	70.3
The FCAT would still be taken seriously if there weren't consequences for students.	32.9	27.4
The FCAT would still be taken seriously if there weren't rewards for teachers.	62.5	11.2
The FCAT is a good indicator of the student's level of mastery for required curriculum.	19.5	44.4
I have control over how my students perform on the FCAT.	22.1	32.3
I spend extra time in class stressing material that is likely to show up on the FCAT.	77.6	11.9

Table 4-2 (continued)

Statement	% Agree	% Disagree
I spend more time going over test taking skills now than before the FCAT was established.	67.7	12.3
The FCAT has an overall positive impact on Florida.	13.5	55.4
The FCAT has an overall positive impact on my school.	17.8	54.8
The FCAT has an overall positive impact on my students.	12.9	61.4
Item types (e.g. multiple choice) used on the FCAT, are the most appropriate type for each learning objective.	21.7	31.8
Prompts used in the writing section of the FCAT are adequate for measuring the student's overall writing ability.	23.9	37.3
Other subjects (e.g. social studies, art) should be included in the FCAT.	20.0	55.7
The FCAT measures the most important concepts in Reading.	50.0	7.8
The FCAT measures the most important concepts in Writing.	34.9	14.2
The FCAT measures the most important concepts in Mathematics.	45.3	12.5
The FCAT measures the most important concepts in Science.	25.0	21.7

Research Question 1

Teachers' opinions of accountability will be significantly different at each of the four areas of accountability: (1) NCLB (2) A+ Plan (3) SSS (4) FCAT. A series of six non-directional dependent sample *t*-tests were performed to test for significant

differences at each level of accountability. The means, standard deviations, and sample sizes for the variables of interest are shown in Table 4-3. A Bonferroni adjustment was made to control for the family-wise Type 1 error rate. The mean difference for SUMNCLB and SUMAPLAN was not statistically significant $t(65) = -.205$, $p = .803$. The mean difference for SUMNCLB and SUMSSS was statistically significant $t(69) = -.9.399$, $p = .000$. The mean difference for SUMNCLB and SUMFCAT was not statistically significant $t(68) = -1.696$, $p = .095$. The mean difference for SUMAPLAN and SUMSSS was statistically significant $t(63) = -8.616$, $p = .000$. The mean difference for SUMAPLAN and SUMFCAT was not statistically significant $t(63) = -.1.127$, $p = .264$. The mean difference for SUMSSS and SUMFCAT was statistically significant $t(67) = 8.417$, $p = .000$. SUMSSS differed significantly from every other group, indicating that teachers rate the SSS differently than the other areas of accountability. All other group differences were not statistically significant.

Table 4-3 Descriptive Statistics for New Variables

Variable	N	Mean	SD
SUMNCLB	72	2.48	0.96
SUMAPLAN	68	2.55	0.96
SUMSSS	70	3.69	0.76
SUMFCAT	69	2.68	0.85
SUMFCATSECT	60	3.30	0.86
AVERAGE	63	2.88	0.66

Research Question 2

GRADE will be a contributing factor to the responses on INSIGHT, IMPACT, SUMNCLB, SUMAPLAN, SUMSSS, and SUMFCAT. Planned complex contrasts were performed to test this hypothesis. The Bonferroni test was used for testing the statistical significance of the simple effects. Schools that received a grade of A or B were combined

and contrasted with schools that received a grade of D. The Bonferroni technique requires family-wise alpha (.05) to be divided by the number of contrasts (six). The means, standard deviations, and sample sized for each variable, broken down by school grade are shown in Table 4-4. The contrast of A and B schools with D schools was statistically significant for INSIGHT, $t(57.92) = 3.716$, $p = .000$. The contrast of A and B schools with D schools was statistically significant for IMPACT, $t(52.35) = 2.985$, $p = .004$. The contrast of A and B schools with D schools was not statistically significant for SUMNCLB, $t(69) = -0.923$, $p = .359$. The contrast of A and B schools with D schools was statistically significant for SUMAPLAN, $t(45.56) = 2.830$, $p = .007$ The contrast of A and B schools with D schools was not statistically significant for SUMSSS, $t(67) = 0.661$, $p = .511$. The contrast of A and B schools with D schools was not statistically significant for SUMFCAT, $t(61.67) = 0.215$, $p = .830$. When appropriate, equal variances were not assumed based on Levene's test for homogeneity of variances.

Table 4-4 Descriptive Statistics for New Variables by GRADE

	A Schools			B Schools			D Schools		
Variable	N	Mean	Std. Dev	N	Mean	Std. Dev	N	Mean	Std. Dev
INSIGHT	25	2.92	1.15	21	1.95	0.74	28	1.61	0.92
IMPACT	25	4.36	0.64	21	3.81	0.98	28	3.46	0.88
SUMNCLB	24	2.74	1.04	21	2.03	0.56	27	2.59	1.03
SUMAPLAN	23	3.30	0.87	18	2.17	0.50	27	2.16	0.89
SUMSSS	23	3.83	1.00	20	3.63	0.51	27	3.61	0.70
SUMFCAT	23	2.91	0.97	19	2.46	0.80	27	2.64	0.75

Research Question 3

YEAR will have a linear relationship with AVERAGE. A simple regression analysis was conducted to examine the degree of association between the outcome variable AVERAGE and the explanatory variable YEAR. The simple model yielded an R^2 of .082 and was statistically significant, $F(1, 57) = 5.076$, $p = .028$, suggesting that

the amount of years teaching (YEAR) is jointly associated with 8.2% of an individuals summated score on the accountability survey (AVERAGE). The adjusted R^2 for the reduced model was .066. Table 4-5 reports the unstandardized regression coefficients (*b*), the standardized regression coefficients (β), the observed *t* statistics, and the squared semi-partial correlations (r^2).

The interpretation of the unstandardized regression coefficient for any explanatory variable is a function of the scale of measurement of that variable. The interpretation of the regression coefficient for a continuous variable can be made in terms of rate and direction of change. The regression coefficient indicates the expected unit change in the outcome variable for each unit change in any explanatory variable, while holding the others constant. For example, YEAR is a continuous variable with an unstandardized regression coefficient of $b = .014$. This suggests that each unit increase in YEAR (i.e., number of years teaching) results in an average .014 unit increase in AVERAGE (i.e., more positive view on modern accountability). Even though an increase of .014 in AVERAGE is statistically significant, it may not be practically significant.

Table 4-5 Summary Statistics for Simple Regression

Variables	*b*	Std Error	β	*t*	*p*	r^2
Intercept	2.635	.124		21.332	.000	
Year	.014	.006	.286	2.253	.028	.082

Research Question 4

Teachers will rate the subsections of the FCAT statistically higher than the FCAT as a whole. A directional dependent samples *t*-test was performed to check for mean differences between SUMFCAT and SUMFCATSECT. The means, standard deviations, and samples sizes of interest are shown in Table 4-3. The difference in the mean response

for SUMFCAT and SUMFCATSECT was statistically significant $t(59) = -7.670$, $p = .000$. Reject the null hypothesis. SUMFCATSECT was rated significantly higher than SUMFCAT.

CHAPTER 5
DISCUSSION

Discussion of Findings

Research Question 1

The average teacher responses for each of the four major areas of accountability (NCLB, A+ Plan, SSS, and FCAT) were compared to one another. The SSS were rated higher than every other area of the accountability design. All other comparisons were not statistically significant, indicating that teachers on average rate NCLB, A+ Plan, and FCAT approximately the same. Teachers may have rated the SSS higher than the others because the SSS are the only accountability topic without direct consequences attached. When teachers lend their opinion, it is probably difficult to separate the area they are assessing and the consequences attached to it. In other words, responses on the quality of the FCAT as a measurement instrument, include the negative feelings towards the consequences. They rate it as a "bad test" independent of the quality of the test because of the negative consequences associated with it. In addition, teachers work with the SSS more closely and most likely know more about SSS than the other areas examined in our study. Teachers appear to have a firmer grasp on the SSS hence rendering their depiction of the SSS more accurate than the areas they have been exposed to less.

Research Question 2

Schools with a grade of A and B were combined and compared to all the D schools on each area of accountability (NCLB, A+ Plan, SSS, and FCAT) in addition to IMPACT (I have an impact on the grade my school receives) and INSIGHT (the grade attached to

each school gives parents insight into how well that school is operating). The purpose of this hypothesis was to test whether teachers from successful schools differed from unsuccessful schools in their opinion on the broad accountability areas, in addition to the two survey items that tapped into how much control a teacher felt they had on their schools grade. Also, we examined if teachers felt the grade accurately depicted their school. The contrasts were statistically significant for INSIGHT, IMPACT, and SUMAPLAN. That is, teachers from schools with differing grades rated the variables associated with school grades different but not other variables associated with different aspects of accountability. Teachers from schools with a low grade rate their ability to impact their schools grade less than teachers from a higher achieving school. Teachers from D schools responded less favorably than the others on the authority of the school grading systems ability to let the public how well that school is actually operating. Underachieving schools also rated the A+ Plan lower than the A and B schools combined, probably because the A+ Plan has a more negative effect on them. This is similar to the previous research question where teacher responses on the quality of the school grading system are affected by the consequences attached. Intuitively, underachieving schools have more negative consequences thus a more negative attitude towards the A+ Plan. All the other contrasts were not statistically significant.

Research Question 3

Teachers were asked how many years they have been teaching. This data was used to test whether there was a linear relationship between years as a teacher and overall attitude toward accountability. There was a small, but statistically significant relationship. The model shows that with each unit increase in years teaching there is a .014 unit increase in overall attitude toward accountability. To illustrate, after 20 years of

teaching, there is only a .28 (on a 5-point Likert scale) increase in overall attitude. This is not practically significant, in that very small and meaningless changes are observed until the difference in time teaching is very large. However, this relationship may have been statistically significant because specific items yielded different responses across teachers with more or less experience. This could be an indicator of awareness by teachers and a thorough understanding of accountability. Perhaps teachers that have been teaching longer understand accountability better and are able to rate it more accurately.

Research Question 4

Opinions held by teachers on the FCAT as a whole were compared to the subsections of the FCAT. Teachers rated the FCAT lower than the sections that comprise it. One possible reason for this is, teachers responded about the consequences of the FCAT instead of the quality of the instrument whereas when responding to individual subject areas are able to focus on just the quality. In other words, there are not direct consequences attached to the FCAT subsections and teachers may be able to answer more fairly. Another possibility for this discrepancy is that teachers do not really know whether or not these are good testing instruments. In lay terms, most teachers rated the FCAT as “bad” and the subsections as “indifferent”. It was not that they praised the FCAT subsections, their opinions were neutral. This could be because they did not know if each section was a good indicator of a student’s level of mastery on a subject area.

Implications of the Descriptive Statistics

In addition to the validity implications of the four research questions addressed in our study, the descriptive statistics contribute greatly to the validity argument. In general, evidence for consequential validity of Florida’s use of the FCAT collected during our study is not favorable. Results state that only 21.7% of teachers surveyed agree that “the

FCAT measures the SSS well", 19.5% of teachers agree that "the FCAT is a good indicator of the student's level of mastery for required curriculum", and 12.9% of teachers agree that "the FCAT has a positive impact on my students".

Teachers were asked to make an evaluative judgment of the quality of the FCAT. As previously stated, results were not favorable. The importance may not be how the teachers rated the FCAT, but why they rated it that way. The approach taken throughout our study was to use teachers as a tool to gauge the validity of the FCAT's use, hence lending part in a validity argument for Florida's entire accountability design. The overall sentiment towards every part of accountability measured in our study, with consequences directly attached, was disapproving or at best, indifferent. It seems likely that the consequences of the FCAT directly contribute to a teacher's evaluation of the quality of the FCAT as a measurement tool.

The results show discrepancies in FCAT ratings. In contrast to the 19.5% of teachers who agreed that the FCAT was a "good indicator of the student's level of mastery for required curriculum", 50%, 34.9%, 45.3%, 25% of teachers agreed that the FCAT measured the most important concepts in reading, writing, mathematics, and science, respectively. This discrepancy could be a result of the text in the items or the propensity of teachers to underrate the quality of the FCAT based on past experiences. Teachers may have negative feelings towards the assessment because they disagree with the consequences attached to it (70.3% of teachers disagreed with "the high-stakes attached to the FCAT are necessary"). Further analysis is needed to draw conclusions on this topic.

Limitations of this Analysis

Response Rate

The response rate was only 28%, usually considered unacceptable. People that respond to surveys have differing characteristics than those who do not return surveys (i.e., respondents tend to be educated and female). Our studies sample was homogeneous (all elementary school teachers, educated, primarily female) perhaps helping to reduce the error (i.e., error may be less than if the sample was more heterogeneous). However, it is possible that the members of the sample who opted not to participate are more apathetic towards accountability issues. The reasons for declining participation are unknown, contributing significantly to the limitations of this analysis.

The response rate also varied across schools. Response rate ranged from 50% through 15.2%. Only five surveys were returned by one of the schools, yielding an unreliable representation of that school. An attempt was made to correct for this error by analyzing similar schools with like responses.

Sampling Issues

One of the schools was from a different district. The implications are mixed. Some results were generalizable across districts. However, the school in central Florida differed on many items to a demographically similar school in north central Florida. Reasons for these deviations are not known. This could be an indicator of the pressures placed on teachers directly from principals or district officials.

Suggestions for Future Research

As mentioned, the issues discussed in our study did not generalize entirely across districts. Further samples from districts and schools across Florida need to be sampled in order to gather more information on the consequences of high-stakes testing. Another

possible angle for future research is examining the factors that can contribute to teacher attitudes on accountability. Our study examined the role of length of time as a teacher and school grade for each respondent on accountability opinions. Additional studies should delve deeper into the demographics, motivations, and backgrounds of their participants.

Notably, 77.6% of respondents indicated that they "spend extra time in class stressing material that is likely to show up on the FCAT". A closer look as to what, according to teachers, constitutes "extra time" and the plausible implications that has on the validity of the assessments interpretation would make for an interesting study.

Closing Remarks

It is hopeful that our study gives insight into some of the effects of high-stakes testing on teachers. The NCLB act places more pressure on teachers than ever before in terms of student achievement. At times, it seems that teachers are being held too accountable for their students learning. Many teachers expressed their frustrations in the margins of the survey explaining that much of what they are held accountable for is beyond their control, and factors like parental support and innate ability are more likely contributing to the child's success, than their efforts. NCLB is still relatively new and only time will tell the exact benefits and repercussions of the new accountability designs.

APPENDIX
ACCOUNTABILITY AND TEACHER ATTITUDES SURVEY INSTRUMENT

I. The following statements address your thoughts on the No Child Left Behind (NCLB) act, Florida's A+ Plan, and Florida's accountability system in general. Please indicate your level of agreement with each statement.

	strongly disagree				strongly agree
1. In general, the Florida accountability system works well…….....	1	2	3	4	5
2. Goals set forth by the NCLB Act will most likely be actualized……..…….....	1	2	3	4	5
3. The NCLB Act has an overall positive impact on the United States…….....1		2	3	4	5
4. The NCLB Act has an overall positive impact on Florida…….....1		2	3	4	5
5. The A+ Plan holds schools accountable for their students learning…….....1		2	3	4	5
6. The grade attached to each school helps give parents insight on how well that school is operating.....	1	2	3	4	5
7. Students from low performing schools should be able to transfer to another school….....1		2	3	4	5
8. The A+ Plan helps motivate teachers and administrators…….....….	1	2	3	4	5
9. I have an impact on the grade my school receives…...1		2	3	4	5
10. Administrators have an impact on the grade their school receives.....	1	2	3	4	5
11. The student body has an impact on the grade their school receives.....1		2	3	4	5

II. The following statements concern the Sunshine State Standards (SSS) and the Florida Comprehensive Assessment Test (FCAT). Please indicate your level of agreement with each statement.

	strongly disagree				strongly agree
12. The SSS adequately outlines the curriculum content at each grade level.....	1	2	3	4	5
13. The SSS lay the foundation for a broad curriculum....	1	2	3	4	5
14. All the SSS will be taught at one point or another.....	1	2	3	4	5
15. The FCAT measures the SSS well.....	1	2	3	4	5
16. The FCAT assesses the most important material at each grade level.....	1	2	3	4	5
17. Florida is enacting the NCLB Act appropriately with the FCAT.....	1	2	3	4	5
18. The high stakes attached to the FCAT are necessary..	1	2	3	4	5
19. The FCAT would still be taken seriously if there weren't consequences for students.....	1	2	3	4	5
20. The FCAT would still be taken seriously if there weren't rewards for teachers.....	1	2	3	4	5
21. The FCAT is a good indicator of the student's level of mastery for required curriculum.....	1	2	3	4	5
22. I have control over how my students perform on the FCAT.....	1	2	3	4	5
23. I spend extra time in class stressing material that is likely to show up on the FCAT.....	1	2	3	4	5
24. I spend more time going over test taking skills now than before the FCAT was established.....	1	2	3	4	5
25. The FCAT has an overall positive impact on Florida..	1	2	3	4	5
26. The FCAT has an overall positive impact on my school.....	1	2	3	4	5
27. The FCAT has an overall positive impact on my students.....	1	2	3	4	5

	strongly disagree				strongly agree
28. Item types (e.g. multiple choice) used on the FCAT, are the most appropriate type for each learning objective.....	1	2	3	4	5
29. Prompts used in the writing section of the FCAT are adequate for measuring the student's overall writing ability.....	1	2	3	4	5
30. Other subjects (e.g. social studies, art) should be included in the FCAT.....	1	2	3	4	5
The FCAT measures the most important concepts in:					
31. Reading.....	1	2	3	4	5
32. Writing.....	1	2	3	4	5
33. Mathematics.....	1	2	3	4	5
34. Science.....	1	2	3	4	5

III. In the following section please tell us about yourself.

What grade level do you teach?_____

If applicable, what subject do you teach?_____

In what school year did you begin teaching? _____

LIST OF REFERENCES

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher,* 33(1), 4-14.

Cizek, G. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19-27.

Crocker, L., Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Wadsworth.

Cronbach, L. (1988) Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test Validity*. (pp. 3-17). Hillsdale, NJ: Erlbaum. (as cited by Haertel, E. (1999) Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice,* 18(4), 5-10.)

Florida Department of Education [FDOE]. (2003). *Consolidated state application Accountability Workbook for State grants under Title IX, Part C, Sec. 9302 for the Elementary and Secondary Education Act (Pub. L. No. 107-110.)* March 26.

Florida Department of Education [FDOE]. (2005a). *2004-2005 School Accountability report.* Last retrieved March 2006. Available online at: http://schoolgrades.fldoe.org/.

Florida Department of Education [FDOE]. (2005b). *Fact Sheet: NCLB and Adequate Yearly Progress*. Last retrieved March 2006. Available online at: http://www.fldoe.org/NCLB/FactSheet-AYP.pdf.

Florida Department of Education [FDOE]. (2005c). *FCAT Web Brochure.* Last retrieved March 2006. Available online at: http://www.firn.edu/doe/sas/fcat/fcatpub1.htm.

Florida Department of Education [FDOE]. (2005d). *Grading Florida Public Schools 2004-2005.* Last retrieved March 2006. Available online at: http://firn.edu/doe/schoolgrades/pdf/schoolgrades.pdf.

Florida Department of Education [FDOE]. (2005e). *Sunshine State Standards*. Last retrieved March 2006. Available online at: http://www.firn.edu/doe/curric/prek12/index.html.

Haertel, E. (1999). Validity arguments for high-stakes testing: In search of evidence. *Educational Measurement: Issues and Practice*, 18(4), 5-10.

Haladyna, T., Downing, S. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice,* 23(1), 17-26.

Hill, R., DePascale, C. (2003). Reliability and No Child Left Behind accountability designs. *Educational Measurement: Issues and Practice,* 22(3), 12-21.

Kohn, A. (2000). Burnt at the high stakes. *Journal of Teacher Education*, 51(4), 315-327.

Lane, S. (2004), Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice,* 23(3), 6-14.

Lord, F., Novick, M. (1968). *Statistical theories of mental test scores*. Reading MA: Addison Wesley. (as cited by Haladyna, T., Downing, S. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice,* 23(1), 17-26.)

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice,* 14(4), 5-8.

No Child Left Behind Act of 2001, Public Law 107-110, §115 Stat.1425, 107th Congress (2002).

U.S. Department of Education [USDOE]. (2006a). *Budget Office- U.S. Department of Education*. Last retrieved on March 2006. Available online at: http://www.ed.gov/about/overview/budget/index.html?src=az.

U.S. Department of Education [USDOE]. (2006b). *No Child Left Behind.* Last retrieved on March 2006. Available online at http://www.ed.gov/nclb/landing.jhtml.

BIOGRAPHICAL SKETCH

Kathryn Miller received a Bachelor of Science degree in psychology from the University of Central Florida (Orlando) in 2003. She enjoyed the quantitative aspect of research psychology and decided to minor in statistics. After graduating, she enrolled as a graduate student at the University of Florida, majoring in research and evaluation methodology in the Department of Educational Psychology. While in graduate school she was fortunate to get the opportunity to be a Graduate Teaching Assistant under Dr. David Miller for the course, Assessment in General and Exceptional Education, where she instructed students on proper assessment procedures. After graduation, Kathryn hopes to relocate to Boston and work on a research team that investigates issues related to health and medicine.