1-10-2014

Using a Computer-adaptive Test Simulation to Investigate Test Coordinators' Perceptions of a High-stakes Computer-based Testing Program

Tiffany Hogan
Georgia State University

ACCEPTANCE

This dissertation, USING A COMPUTER-ADAPTIVE TEST SIMULATION TO INVESTIGATE TEST COORDINATORS' PERCEPTIONS OF A HIGH-STAKES COMPUTER-BASED TESTING PROGRAM, by TIFFANY HOGAN, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Philosophy, in the College of Education, Georgia State University.

The dissertation Advisory Committee and the student's Department Chairperson, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
T. Chris Oshima, Ph.D.
Committee Chair

_____
Janice Fournillier, Ph.D.
Committee Member

_____
Kijua Sanders-McMurtry, Ph.D.
Committee Member

_____
William Curlette, Ph.D.
Committee Member

_____
Date

_____
William Curlette, Ph.D.
Chairperson, Department of Educational Policy Studies

_____
Paul A. Alberto, Ph.D.
Interim Dean
College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's Director of Graduate Studies, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Tiffany Elaine Hogan

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is

Tiffany Elaine Hogan
3845 Brookview Point
Decatur, Georgia 30034

The director of this dissertation is

Dr. T. Chris Oshima
Department of Educational Policy Studies
College of Education
Georgia State University

VITA

Tiffany Elaine Hogan

ADDRESS: 3845 Brookview Point
Decatur, Georgia 30034

EDUCATION:

Ph.D.	2013	Georgia State University Educational Policy Studies
Ed.S.	2001	Georgia State University Science Education
L-5	2000	University of Georgia Leadership Certificate
M.Ed.	1999	Georgia State University Science Education
B.S.	1993	Spelman College Natural Science/Biology

PROFESSIONAL EXPERIENCE:

2012-present	Race to the Top Psychometrician Clayton County Public Schools, Jonesboro, GA
2005-2012	Assistant Principal East Clayton Elementary, Ellenwood, GA
2004-2005	Assistant Principal/Career Technical Supervisor Riverdale High School, Riverdale, GA
2003-2004	Assistant Principal Forest Park High School, Forest Park, GA
2002-2003	Instructional Technology Specialist Forest Park High School, Forest Park, GA
1993-2002	Science Teacher Dekalb County Public Schools, Decatur, GA

PROFESSIONAL SOCIETIES AND ORGANIZATIONS:

2009-2010	American Educational Research Association
2008-2009	Kappa Delta Pi
2008-2009	Sisters of the Academy
2005-2006	American Educational Research Association

PRESENTATIONS AND PUBLICATIONS:

Hogan, T. (2006, April). The ebb and flow of educational leadership. Paper presented at the Sources in Urban Educational Excellence Conference, Atlanta, GA.

ABSTRACT

USING A COMPUTER-ADAPTIVE TEST SIMULATION TO INVESTIGATE TEST COORDINATORS' PERCEPTIONS OF A HIGH-STAKES COMPUTER-BASED TESTING PROGRAM

by

Tiffany E. Hogan

This case study examined the efficiency and precision of computer classification and adaptive testing to elicit responses from test coordinators on implementing high-stakes computer-based testing. Test coordinators from five elementary schools located in a Georgia school district participated in the study. The school district administered state-made, high-stakes tests using paper and pencil; locally-developed tests via the computer or paper and pencil. A post-hoc simulation program, Comprehensive Simulation of Computerized Adaptive Testing, used 586 student item responses to produce results with a variable termination point and a classification termination point. Results from the simulation were analyzed and used in the case study to elicit interview responses from test coordinators. The photographs of computer-labs and test schedule documents were collected and analyzed to validate school test coordinators' responses.

Test coordinators responded positively to the efficiency and precision of simulation results. Some test coordinators preferred the use of computer-adaptive tests for diagnostic purposes only. Test coordinators' experiences focused on the security, the emotions, and the management of testing. The findings of this study will benefit those interested in implementing a high-stakes, computer-based testing program by recommending a simulation study be conducted and feedback be solicited from test coordinators prior to an operational test administration.

USING A COMPUTER-ADAPTIVE TEST SIMULATION TO INVESTIGATE TEST COORDINATORS' PERCEPTIONS OF A HIGH-STAKES COMPUTER-BASED TESTING PROGRAM

by

Tiffany E. Hogan

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Educational Policy Studies
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, Georgia
2013

ACKNOWLEDGMENTS

> And those he predestined, he also called; those he called, he also justified; those he justified, he also glorified. What, then, shall we say in response to this? If God is for us, who can be against us? Romans 8: 30-31

This dissertation was a journey of love. Without love, nothing is possible. It is with Love that I acknowledge all of the wonderful people who helped me on this journey. These individuals played various roles in my journey. Some were on the sideline cheering me on; some offered their expertise and knowledge, whereas others offered a sympathetic ear, while others allowed me the flexibility to complete my journey. Without these individuals, none of this would have been possible.

To Dr. Oshima, my committee chair, the reason why I pursued a dissertation on computer-based testing. Your knowledge and expertise of Item Response Theory is equal to none. This dissertation would not have come to fruition without you. Also, thank you for suggesting that I review the Computer-Adaptive Testing Simulation Program. Thank you, for agreeing to chair my committee. I will forever be grateful for your kindness.

To Dr. Nathan Thompson (Assessment Systems Corporation), albeit you do not know me, your work with computer-adaptive testing has been nevertheless inspiring. Thank you for your kindness and assistance when I could not figure out the problem with running my data set. Thank you for the use of the CATSim Program for my dissertation.

To Dr. Fournillier, my committee member, I thank you for your advice and wisdom. Your knowledge and expertise in qualitative research is astounding. To Dr. Curlette, committee member, thank you for so graciously agreeing to serve on my committee and for giving me sound advice regarding my prospectus. To Dr. Sanders-

McMurtry, committee member, thank you for giving me direction and outstretching your hand to lift me up. To my family, friends, and co-workers, I could not have made it on this journey without your support. To my husband, Jonathan Hogan, you have been by my side this entire journey, some days you were on the sideline cheering me on and other days you had to run alongside me. I thank you for your enduring love and support.

To my friends, Dr. Yolanda Troutman and Dr. Machelle Matthews, thank you for sharing your knowledge and expertise as well as being an inspiration for me. Thank you to the teachers and staff at East Clayton Elementary who cheered me on for seven years of my journey. Thank you, to my current supervisor, Dr. Delphia Young, for being an inspiration to me in completing my journey. Thank you, Mrs. Marcia McAllister, one of the last great English teachers, for your dedication and support. Thank you to the countless numbers who remain nameless, for I see your face and know that you too were an integral part to my success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABBREVIATIONS

CAT	Computer Adaptive-Tests
CATSim	Computer Adaptive Test Simulation
CBT	Computer Based Test
CCT	Computerized Classification Test
CI	Confidence Interval
CTT	Classical Test Theory
DOE	Department of Education
ICC	Item Characteristic Curve
IRT	Item Response Theory
LEA	Local Educational Agency
MLE	Maximum Likelihood Estimation
NCLB	No Child Left Behind
PARCC	Partnership for Assessment of Readiness for College and Career
PPT	Paper and Pencil Tests
RMSD	Root Mean Square Difference
RTTT	Race to the Top
SBAC	Smarter Balanced Assessment Consortium
SE	Standard Error
SEM	Standard Error of Measure
USED	United States Education Department
VL-CAT	Variable Length Computer Adaptive Test

CHAPTER 1

THE PROBLEM

Technology has become a dominant force for engaging students in the classroom. From Smartboards, to iPods, to classroom blogging, students today are inundated with technology to enhance their learning. The convergence of technology and its use in the classroom seemed almost effortless. Students easily grasped at the use of technology devices in the classroom. Now that technology has infiltrated school districts at the instructional level, the next step is for it to become more systemic. It is meant for technology to become a mechanism for high-stakes testing.

The ease at which technology and classroom instruction merged was not of much debate. Actually, the emergence of instructional technology was a welcome change to the day to day repetition of classroom instruction. Technology as it relates to high-stakes testing was unfortunately, the complete opposite. The use of computers for high-stakes testing has been received with mixed reviews. Test developers and policymakers view the use of computers for high-stakes testing differently, often times debating its feasibility. This much-contested debate has now become a reality for school districts. By 2015, school districts will have to administer high-stakes tests via the computer. School administrators are currently faced with transitioning teachers and students from paper and pencil tests (PPT) to computer-based tests (CBT).

Although, CBTs are the next wave of educational assessments, its implementation will come with both benefits and challenges. Among the benefits touted by technology and assessment experts are data quality, score reporting, logistics and low administration cost, whereas challenges include infrastructure and scheduling (Grunwald Assoc., 2010).

There are still additional challenges which persist, such as cost, resources, and knowledge (Trotter, 2003). Virginia's Department of Education began administering CBTs during the 2000-2001 school year. State officials noted that initial implementation was not easy, with the lack of computer resources being the main problem. One important suggestion made by Virginia was to have individuals who were knowledgeable of the technology. It became evident to Virginia officials "that there were not separate technology issues and separate assessment issues . . . if you have one you have to have the other" (Grunwald Associates, 2010, p.7). As school districts prepare for the implementation of CBTs, addressing these challenges becomes an essential part to the fidelity of the testing program.

There are many benefits to CBT, interactive screens, adaptive testing, and electronic scoring to name a few. Unfortunately, the implications for implementing a large-scale CBT program are far reaching (Davey, 2011). In order for states, as well as school systems, to have success, there is a need for a detailed plan of transition. The plan should include a comparison of new and old assessments, cost of new assessments, and professional development (Achieve, 2010). Technology infrastructure, the number of computers, and the length of the testing window are all interrelated issues that have to be addressed prior to CBT becoming operational (State Educational Technology Directors Association, 2011).

When technology merges with assessments, a different product emerges offering a new form of design, a new mode of administration, and a new form of score reporting. Thus, the merging of technology and assessments offer an easier and more efficient way to meet the requirements of NCLB through the delivery, score reporting, and results

analysis of CBTs. The confluence of testing and technology is ideal for educational accountability systems such as NCLB. Though this merging shows promise for accountability systems, one of its biggest impediments is cost. Money for education has been limited in recent years. Ensuing budget cuts at the state and local levels along with increased testing in primary and secondary schools has slowed the implementation of CBTs (Olson, 2003).

Problem Statement

My experience as a former school test coordinator was the impetus for this study. As a doctoral student, studying educational measurement, I understood the importance of conducting simulation studies prior to the administration of a computer-based test. As a former test coordinator, I also understood the challenges of coordinating a high-stakes test. From my experience, coordinating a high-stakes test took an insurmountable amount of time, patience, and organization. I can vividly recount the numerous test booklets that I so meticulously logged the names for each and every student on the security checklist. I remember the stacks on stacks of boxes that so punctually arrived the week prior to spring break to be inventoried; the scheduling teachers as a test administrator or proctor;and the daily counting of test booklets and answer documents. .

It was from these memories that I welcomed the idea of computer-based testing. I remember thinking that with computer testing, the daily counting and documenting of test booklets and security issues related to erasures, lost answer documents, test booklets would all become non-existent. The only problem that I could foresee with computer-based testing was the limited number of computers. I even had a solution for this . . . a computer-adaptive test that would terminate once a student reached a certain level of

mastery or precision. Yes, computer-adaptive testing was an efficient and precise way to measure student learning.

Research Questions

To overcome the challenges associated with high-stakes computer-based testing by the year 2015, it is pertinent to gain the perspectives of all stakeholders -- policymakers, test developers, students, and school test coordinators. This study examined the following questions: (1) Are the efficiency and precision of computer-adaptive tests and computer classification tests equivalent or superior to those of paper and pencil benchmark tests? (2) What are test coordinators' perceptions of administering high-stakes tests on the computer? (3) To what extent, if any, did computer simulation results elicit a change in test coordinators' perceptions of administering high-stakes tests on the computer?

Problem Background

A new wave of testing is on the horizon. Gone are the days of "bubble" sheets and erasure marks, counting answer documents and test booklets, packing and re-packing boxes on top of boxes of "top secret" test materials. The new millennium has brought a new form of testing . . . computer-based tests. In March of 2010, President Obama challenged the U.S. Department of Education Office of Technology to develop a National Educational Technology Plan (NETP). This plan called for a "revolutionary transformation" of the technology system in the areas of teaching and learning, assessments, infrastructure, and productivity. As a result, the assessment component required the use of technology-based formative and summative assessments to be used for diagnostic and accountability purposes in educational systems (U.S. Department of

Education, 2010).

In conjunction with President Obama's challenge to NETP, the Secretary of Education pledged a $350 million grant as part of the Race to the Top Initiative (RTTT), for the development of computer-based assessments aligned to the Common Core State Standards. As a result, two consortia, Partnership for the Assessment of Readiness for College Careers Consortium (PARCC) and Smarter Balanced Assessment Consortium (SBAC) were awarded the $350 million grant. Upon accepting the assessment grants in 2010, states are required to implement the grant by the 2014-2015 school year in grades 3-12 (U.S. Department of Education, 2010).

The context of this case study was a school system located in the southeastern region of the United States. Geographically, the school system is found within a county that is bounded by five other counties and near a major city. The school system is often characterized as an urban district due to its demographics, its classification as a metropolitan county, and its proximity to the downtown city limits.

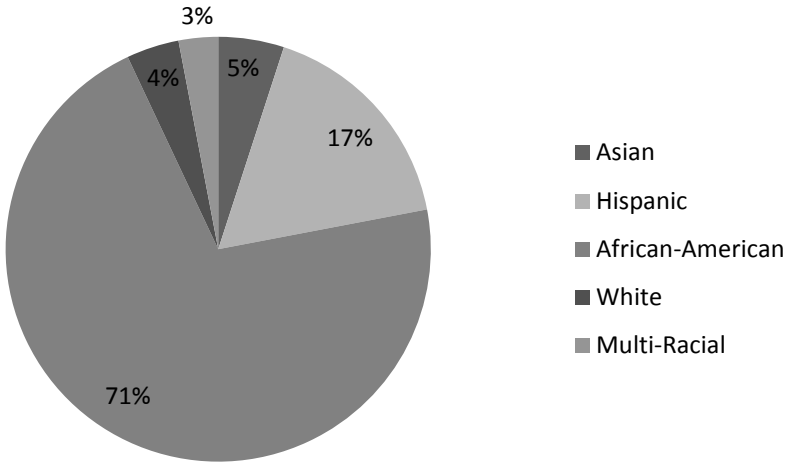


Figure 1. Student Demographics

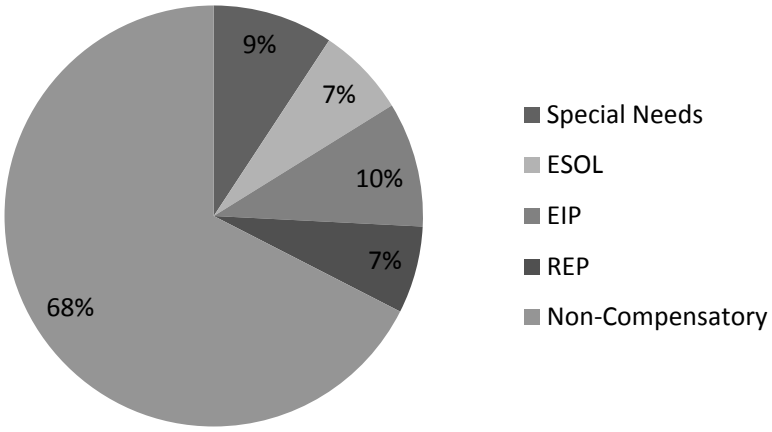


Figure 2. Compensatory Programs

Figure 1 displays the school districts demographics, which consists of students from a variety of ethnic backgrounds; Asian-5%, Hispanic-17%, African-American-71%, White-4%, and Multi-Racial-3%. The school system offers a range of academic programs for students which include; Gifted, Career and Technical, College Preparatory, and International Baccalaureate. Figure 2 shows the percentage of students in

Compensatory Programs: Special Needs-9.3%, English to Speakers of Other Languages-6.9%, Early Intervention Program for grades 3 through 5--9.6%, and Remedial Education for grades 6 through12-6.7%. The school system receives Title 1 funds due to more than 75% of the students receiving free or reduced lunch (Governor's Office of Student Achievement, 2011).

Students are exposed to a wide-range of assessment measures that are both formative and summative. Formative assessments are on-going assessments that are administered to students throughout the school year. These assessments differ in test length, frequency of assessment, and purpose of the assessment. Assessments that are administered 2-3 times a year are used to evaluate student performance as compared to other students in the district at specific intervals. Other assessments are administered once a unit has been completed. This assessment allows the teacher to determine student mastery, as well as, monitor teacher progress on the pacing chart. Another type of formative assessment consists of 5-10 test items and provides teachers with immediate feedback on student understanding of a topic.

Like formative assessments, summative assessments are administered in grades kindergarten through twelfth. Summative assessments are standardized tests that are administered statewide and in some instances nationally. There are multiple summative assessments that are administered to students for a myriad of reasons. Some summative assessments are administered at the end of a grading period or the conclusion of the school year such as end of course tests and graduation exit exams. Other summative assessments are administered to assess students with disabilities, to assess student's writing ability or to assess English Language Learners. National tests are administered to

students at specific grade levels to assess how those students compare to students at that grade level across the nation.

Significance

The purpose of this study was to investigate if the efficiency and precision of computer adaptive and computer classification tests were equivalent or superior to paper and pencil benchmark tests and to explore how the simulation results changed the perceptions of school test coordinators on high-stakes computer-based testing. The importance of this study is due to the limited amount of research on high-stakes computer-based testing for K-12 education. The perspective of test coordinators will provide educational leaders and policy makers with the possible challenges as well as solutions to implementing a high-stakes computer-based testing program. Another significant point to this study was the discourse elicited from the results of the computer simulation. Test coordinators' perceptions of simulation results may prove to be noteworthy because of their practical insight to implementing a computer-based testing program.

Limitations

The limitations of this study were:

1. The number of examinee's responses used for the simulation. Due to the size of the school district it was necessary to take a sample of the population.
2. The theta cutoff level for the Computerized Classification Test was determined by the researcher.
3. The examinee's ability to read and understand the subject matter of the test.

4. The number of items administered to each examinee.

5. The responses from examinees that received testing accommodations.

6. There is a potential for bias because the researcher is also a test coordinator in the school district. The researcher also knows personally some of the participants in the study.

Assumptions

The assumptions in this study were as follows:

1. The Computer Adaptive Test Simulation (CATSim) Program was a valid and reliable simulation program.

2. Examinee responded to test items to the best of their ability.

3. Examinees and test coordinators were a representative sample of the population.

4. School test coordinators responded to interview questions to the best of their ability.

5. The benchmark assessment was a valid and reliable test.

CHAPTER 2

REVIEW OF LITERATURE

Chapter two reviews pertinent literature as it relates to the research questions posed in this study. This chapter begins with a description of terminology and follows with supporting literature. The purpose of the next section was to provide the reader with an overview of theory and terminology as it relates to the following research question: Are the efficiency and precision of paper and pencil benchmark tests equivalent to those of computer adaptive tests and computer classification tests? A basic understanding of relevant terminology will also prove useful in chapter four, in the analysis of simulation results.

Terminology Description

There are two main frameworks in testing theory: item response and classical test. In Classical Test Theory (CTT) an examinee's ability level is determined by the concept of a true score or the estimated score of an examinee on a specific test. Item response theory (IRT), however, relies on the notion that examinee test performance can be predicted by "traits, latent traits, or abilities" (Hambleton, Swaminathan, & Rogers, p.7). As a result, IRT characterizes the relationship between an examinee's item performance and traits by a "monotonically increasing function called the Item Characteristic Curve (ICC) (Hambleton et al., p.7)." The ICC shows that as an examinee's ability level increases, the probability of an examinee's correct response will also increase.

The IRT framework for the current study applied the use of the one-parameter, Rasch Model. This model implies a location parameter, the *b* parameter, which measures the difficulty level of the item. The *b* parameter shifts along the ability scale to the left or

right as the difficulty level of the item increases or decreases (Hambleton, et al., 1991). The one-parameter model was applied because the population size (N) required to conduct the study was between 250 and 500 (Jones, Smith, & Talley, 2006).

The premise of IRT resides in the following claims: examinee item responses can be predicted based upon their ability level and the relationship between ability level and item response are characterized by an item characteristic curve. IRT proclaims several assumptions: unidimensionality, local independence, and the relationship between examinee ability level and item responses are represented in the ICC. Unidimensionality is the assumption that only one ability is measured for each item. Local independence assumes that examinee's responses are independent of each other when ability levels are held constant (Hambleton, et al., 1991).

These principles are more easily understood when compared to CTT. For example, true score and observed scores of examinees are the foundation of CTT, whereas ability scores are the foundation of IRT. The feature that separates the two is test dependency. In the case of CTT, true scores are test dependent, whereas in the case of IRT, true scores are test independent. For instance, the true scores on a difficult test would be low, whereas true scores on an easy test would be high. This, however, is not the case for ability scores that are associated with IRT. Ability scores are found to remain constant regardless of the difficulty level of the tests (Hambleton, et al., 1991).

There are many advantages to using an IRT framework. There are different forms of CBTs. Some CBTs are linear or fixed form and some are adaptive. Linear- CBTs are synonymous to PPTs. Each examinee is administered the same number of items in the same order with the only difference being the mode of test administration, computer or

paper and pencil (Davey & Pitoniak, 2006). Linear-CBTs are the least complex and lowest in cost compared to other CBTs. In addition to low complexity and cost, linear-CBTs level of precision is equal to that of PPTs (Davey, 2011).

Computer-adaptive test (CAT) is a type of CBT. A distinguishing feature of CAT compared to linear-CBTs is how "it adjusts the difficulty level of the items so that the examinee's scores best reflects the examinee's ability" (Impara & Foster, p. 111, 2006). Other unique features of CAT are within the termination criteria——fixed and variable-length. Termination criteria are points set by the test developer for the test to end. Fixed-length CAT terminates at a fixed number of items for all examinees, whereas the termination point for Variable Length (VL)-CAT depends upon the examinee (Weiss & Guyer, 2010). Test precision is increased in VL-CAT by manipulating the stopping rule for theta estimates and the standard error (SE) of theta estimates (Impara & Foster, 2006).

The SE of theta estimates is defined by the notion that "the amount of information provided by a test at theta (θ) is inversely related to the precision with which ability is estimated at that point (Hambleton, Swaminathan, & Rogers, 1991, p.94)". The size of SE of theta estimates is inversely related to test length. For instance, the more items that are on a test the smaller the SE of theta estimates. When maximum likelihood estimates (MLE) of θ estimates are obtained, as in the case of CAT, the SE of theta estimates distributions are normal in tests with 10 or greater items (Hambleton, et al., 1991).

The current study employed the use of MLE to estimate an examinees ability level. MLE uses examinee item response patterns to determine item parameters, which subsequently estimates the ability level of the examinee. The MLE method is an essential part of computer-based testing. Specifically, the MLE procedure used in CAT

is to estimate examinees ability level in order to generate test information at the estimated ability level. The test information along with the SE of theta estimates is a necessary part of the administration of the next test item to examinees in computer-adaptive testing (Hambleton, Swaminathan, Rogers, 1991).

A pertinent part of this study examined the efficiency and precision of VL-CAT and CCT through set termination points of .005 or less successive change in standard errors of theta estimates and the theta estimates plus or minus 2.00 standard errors above or below the theta cutoff value of 1.00 (Weiss & Guyer, 2010). The use of a termination point for all examinees allowed for an equal level of measurement precision. Classification termination points established by using Confidence Intervals (CI) at specified theta levels are referred to as Computerized Classification Testing (CCT). CCT terminates when the standard error of theta estimate is plus or minus the CI above or below the cutoff value of 1.00. CCT classifies examinees into categories of either "Pass" or "Fail". Classification is based upon the following criteria: if the theta estimate plus or minus the CI fall above the cutoff value of 1.00, then the examinee "passed", if it falls below the cutoff value of 1.00, the examinee "failed", and if the CI contains the cutoff value of 1.0 then another item is administered to the examinee (Weiss & Guyer, 2010).

VL-CAT and CCT are an essential part to understanding the efficiency and precision of CAT. For the purposes of this study, test efficiency is defined in terms of test length. For example, a shorter test is more efficient than a longer test because the shorter test requires less time to administer. Test precision is defined as accuracy in estimating ability levels. Simulation studies allow for the manipulation of test length needed for an appropriate level of test precision (Davey & Pitoniak, 2006). Simulation

studies are a more cost efficient and convenient way to determine if the test is suitable for a group of examinees. Item response theory is key to simulating a test administration by the use of ability levels and item parameters. By using simulations, a variety of testing environments can be set by manipulating termination points, thus, precision and test length (Parshall, et al., 2002).

This study employed the use of the CATSim Program to evaluate the efficiency and precision of CAT using different termination points. Two different simulations were conducted using examinee item responses. This simulation, referred to as a post-hoc, used real data to create an item response matrix. CATSim results were used to measure efficiency and precision by comparing CAT theta estimates to that of the full item bank. CAT efficiency was measured by the number of simulated items resulting from each test; precision was measured by using Pearson (r) correlation and the Root Mean Square Difference (RMSD) between the CAT theta and thetas of the full item bank. RMSD measures the difference between CAT's θ estimates and true theta using the following equation:

$$\text{RMSD} = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}j - \theta j)^2}{N}}$$

Paper and Pencil

If an individual were asked, "Did you take your test on the computer or with paper and pencil?" More times than not, the answer would be paper and pencil. Most large scale testing programs use paper and pencil for students to code the response of either a multiple-choice, true/false, and/or matching test (Cohen & Wollack, 2006). Multiple-choice tests had been the pre-dominant form of large-scale paper and pencil test up until

the late 1980's when interest in performance-based tests arose. However, this interest quickly waned due to various reasons, and multiple choice tests were once again placed at the forefront. Suggested reasons for the re-emergence of multiple-choice test were related to the ease of use in large-scale testing and cost (Koretz & Hamilton, 2006).

Although paper and pencil tests are a widely accepted form of test administration, it does not mean it is necessarily the most efficient process. In order for a test to be operational there are a multitude of steps as well as individuals who are part of the process. The process to making a test operational may start 12-15 months prior to the administration. Parts of the process range from scheduling test dates to preparing the test site for administration. Test administrators have to be screened, test materials have to be inventoried, and test coordinators have to attend training (McCallin, 2006).

Detailed steps required after administration are scanning and processing. Technology plays a key role in each of these steps. Scanning machines are used to convert paper and pencil responses to computer form. Large-scale testing programs use scanners, known as optical mark readers (OMR), to scan thousands of answer documents within an hour. As part of the scanning process, answer documents have to be examined to ensure feed through to the machine, scanning calibrations, and checking for answer document errors. Once answer documents have been scanned and processed they are ready for scoring (Cohen & Wollack, 2006).

A challenge with administering PPT is ensuring test validity. In an effort to maintain the validity of PPT, it is necessary to make sure the integrity of the test is not compromised. Test integrity is compromised when the security of the test is put at risk due to repeated use of a test item, unauthorized use of resources on the test, erasing

student answers, lost test booklets or answer documents, and allowing extra time. Breaches in test security can take place by the teacher, student, or administrator. In response to compromised test security, several measures are in place to detect such security violations---copying indices, score gain and pattern algorithms, and erasure analysis (Cohen &Wollack, 2006).

There were an insurmountable number of articles comparing the use of PPT and CBT. In 2008, the Texas Education Agency (TEA) provided a technical report on the comparability of scores from PPT and CBT. This report examined the need for comparability studies, methods of data collection, and level of analysis in which the scores are compared. TEA's findings showed the scores for PPT and CBT to be comparable, however examinees were more likely to score higher on Constructed Responses (CR), as opposed to Multiple Choice. Mode effects were found in the Texas Assessment of Knowledge (TAK) Program due to content that required examinees to scroll through a large amount of text. In summary, comparing modes of test administration in the forms of PPT and CBT were similar. These modes of administration are said to be linear due to scoring procedures and test items being the same (Texas Educational Agency, 2008).

Two major concerns of paper and pencil assessments are in regards to the diagnostic information and reporting. In a Minnesota school district, administrators and teachers reported that the statewide high stakes test did not give enough diagnostic information on students. High-stakes test lacked discrimination on students' strengths and weaknesses. Another confounding problem was the extensive amount of time it took for the district to receive the score reports. For

example, schools that administered tests in March and April did not receive score reports until September. Although the monetary costs of testing were not discussed, administrators and teachers voiced concern over the cost of the test with regards to instructional time as compared to the information they received from the results. There were also additional areas of concern regarding the assessment, such as difficulty level and measuring student growth. Teachers and administrators viewed the test as having the same difficulty level for all students to be a point of contention for students with a low ability level. They also viewed the test as lacking the ability to measure individual student growth or growth within a cohort of students (Yeh, 2006).

Computer Based Test

There are different modes of test administration, which examinees receive. The two predominate modes are computer-based and paper and pencil. Linear-CBT and PPT are fixed form assessments. The only difference in a fixed form computer test and a fixed form PPT is the mode in which the test is administered—computer or paper and pencil (Kolen & Brennen, 2004). There have been many studies conducted on the use of CBT versus PPT as it relates to student high stakes testing. However, the conclusiveness of the studies is still a topic of much debate.

Computer-based tests are tests administered on the computer. In some instances, these tests are quite simple in design. For example, linear-CBTs are PPTs that are administered on the computer. Other types of computer-based tests are computer-randomized tests, which randomly administer items from a test bank to examinees (Davey & Pitoniak, 2006; Kolen & Brennan, 2004). However, there is still another type

of computer-based tests that is more complex due to its ability to administer items that are specific to an examinees ability level. This type of computer-based tests is referred to as a computer adaptive test (Kolen & Brennan, 2004).

Since the infusion of technology, CBTs have become a desirable mode of test administration due to its effectiveness in cost, security, score reporting, and its ability to test students continually (Parshall, et al., 2002; Wise & Plake, 1989). Regardless of the many attributes of CBTs as opposed to PPTs, uncertainty exists on the impact of mode of administration on examinee responses (Poggio, et. al., 2005). One reason this debate continues is due to States' Department of Education who implement both CBTs and PPTs. When CBTs and PPTs are both offered within a state, it is difficult to compare test scores from tests that were administered differently.

As previously discussed, computers have a significant role in the processing of score reports for paper and pencil test. Answer documents are scanned, and the uploaded information is then processed via the computer. Today, computers are taking a more prominent role in the test administration process. Computers are now used as a form of test administration instead of paper and pencil.

Unlike PPTs, CBTs do not require the use of a scanner and processing to produce scores. More importantly, however, with CBT certain steps are required prior to the administration of the test. This includes but is not limited to software, hardware, and/or internet bandwidth. Test security is not unique to PPT. Ensuring the integrity of a CBT is also necessary. Compromises in test security are different in CBT due to the uniqueness in administration, as is the case for scheduling examinees to use a computer. This poses an increased threat to security because

multiple examinees may be tested in one day; testing windows are extended, thus, increasing the item exposure (Parshall, et al., 2002).

Parshall et al. (2002), listed several confounding factors examinees had regarding CBTs. Some of these concerns are lack of prior computer experience and the ease of software use. Proposed solutions to these concerns were providing examinees with computer use information prior to test administration and receive interface feedback from the target audience prior to the administration of the test.

The process used to compare alternate forms of a test is known as equating. "Equating is a statistical process that is used to adjust scores on test forms so that scores can be used interchangeably" (Kolen & Brennan, 2004, p. 2); thus before comparing scores of PPT to linear-CBT and CAT test score equating must take place (Kolen & Brennan, 2004). There have been mixed reviews from the outcomes of comparability studies of online assessments and paper and pencil. However, conducting comparability studies poses a unique set of challenges. Researchers, as well as psychometricians, agree there is a need for more controlled experimental comparability studies; unfortunately, this is not a feasible reality when using real data from statewide testing program.

In 2007, Way, Um, Lin, & McClarty, conducted a comparability study using a matched samples analysis. The study used covariates comparing computer-based to paper and pencil test. Test scores were equated, and a score conversion table was used to control for mode effects. A bootstrap approach was used to create raw to scale score conversions by equating online scores to paper and pencil. By using this approach a sample of student computer test scores were matched to student scores from paper and pencil test. A statewide eighth grade test was used to simulate data for 60,000 simulees.

As predicted, there were differences in modes of test administration. The Matched Samples Comparability study displayed mode effects only when the data were simulated. This method worked best when ability groups were equally matched for computer-based and paper and pencil test.

A second area of contention is the interchangeability of scores from CBTs and PPTs. There are mixed results in studies that examine score equivalency as it relates to CBTs and PPTs (Mazzeo and Harvey, 1988). Equivalency guidelines set forth by Computer-Based Test and Interpretations of the American Psychological Association (1985) stated, "the test developer is responsible for ensuring that equivalent results are obtained with the two versions when a computerized version of a paper-and pencil test is constructed" (1985 as cited in de Beer & Visser). In 2005, a quasi-experimental study was conducted evaluating the impact of mode of administration on seventh grade student's math scores. Approximately 644 students were randomly assigned to both modes of administration. The result of the study showed no statistically significant difference between CBT and PPT (Poggio, et al., 2005).

In 2008, the Texas Education Agency (TEA) Technical Report cited the reason for offering CBTs for their testing program was due to the "greater flexibility in administration, reduced administration burdens on district personnel, and the possibility of faster score reporting" offered by CBTs (Texas Education Agency, 2008, p.6). Educational leaders in Virginia touted an increase in efficiency and precision of data collection as a benefit. Although Virginia has implemented high-stakes online testing for the past 10 years, the majority of their tests are paper and pencil because of infrastructure logistics (Schaffhauser, 2011).

The General Scholastic Aptitude Test (GSAT) was administered to South African high school students in a 1998 comparability of study PPTs, linear-CBTs, and CATs. The overall purpose of the study was to compare the results of PPTs and CATs in an effort to make adjustments to the CATs. The study administered both PPTs and CATs versions of the tests to 242 students. The results found the scores on the PPTs to be higher than those of the linear-CBTs as well as the CATs. An examinee unfamiliar with linear-CBTs and CATs was listed as a reason for the non-equivalence of test score results (DeBeer, M. & Visser, D., 1998). Despite opposition and security challenges that face CBTs, the administration of online assessments is the wave of the future. Testing companies are now making the shift to CBTs by offering a variety of formative and summative assessments online (CTB-McGraw-Hill, 2013).

As accountability systems change, the types and number of assessments administered to students also changes. Assessments known as Student Learning Objectives are a way of measuring student growth through the administration of pre- and post-assessments. Student Learning Objectives are administered in grades K-12 for all subjects that are not tested by a state administered summative assessment. In some instances, there could be over 300 Student Learning Objectives Assessments administered within a given school year. A district leader had the following comment regarding Student Learning Objective Assessments "Due to the number of tests administered, we need to be able to administer Student Learning Objective Assessments on the computer. The number of items administered to students should be minimum as to give a snapshot of what the student knows or needs to learn . . . maybe 10-15 items."

Computer Adaptive Test

Forms of adaptive testing have appeared since the early 1900's. However, it was not until Fred Lord of the Educational Testing Service (ETS) seriously began to conduct research in the area of adaptive testing that it began to take form. It was Lord's desire to create a test that efficiently measured the ability levels of both high and low level examinees. According to Lord, theoretically adaptive test would shorten the length of tests without the loss of measurement precision by the administration of test items that would maximize information about an examinees ability level. Adaptive testing only became a reality with the introduction of the computer—thus the term CAT (Hambleton, et al., 1991).

Today, the use of CAT for high stakes test is an area of much contention as well as debate. CAT offers an array of benefits to all stakeholders; each student receives a test adapted to their ability level, test results are immediate, numerous reports can be generated, and large item pools can provide multiple test administrations. One point of contention still remains—grade level testing. For example, opponents of CAT argue that 4th grade students who are not on grade level would be administered questions that are below their grade levels. This argument also stands true for students who perform above grade level (Horn, 2003). As discussed previously, the requirements of NCLB hinder states from using computerized-adaptive testing for high stakes test.

In 2004, Kingsbury and Hauser presented a paper at the annual meeting of the American Educational Research Association entitled, *Computerized Adaptive Testing and No Child Left Behind*. This paper provides evidence of why computerized-adaptive testing is not only an efficient way for meeting the requirements of NCLB, but also an

effective way to determine student ability (Kingsbury & Hauser, 2004). The following uses of test scores have been outlined by *NCLB*: use of proficiency categories for accountability, achievement growth, and to inform instruction. Hauser and Kingsbury compared the utility of linear-CBTs to CATs as it relates to meeting the requirements of *NCLB*. The Rasch model (Item response theory, one-parameter logistic model) was used to calibrate all tests and items. Four sets of linear tests were used with difficulty levels at the 35th and 70th percentile in both math and reading. The linear tests had item difficulties within the following ranges: 36 percent between mean and 1 standard deviation (SD), 9 percent between 1 and 2 SD, and 5 percent between 2 and 3 SD. Reading and math CAT scores were retrieved from the spring 2003 administration. The study compared the amount of information produced from each test as a result of a range of scores. The study concluded that CATs provided more information about student ability than linear test. The percent of students who were not measured with precision was less than 1 percent for CATs and greater than 6 percent for fixed form test. Hauser and Kingsbury contend that the use of CAT to meet the requirements of *NCLB* will allow for the following: challenging questions for students without frustration, accurate scores, and efficiency in score reporting.

A key obstacle in instituting a CAT is the misnomer that it is grade level testing. In 2010, David Harmon, a program specialist for the United States Department of Education stated that,

- Individual level assessments (adaptive assessments) would measure the performance of some students at a particular grade level against lower standards.
- This would result in some schools being held to lower standards than other schools.
- Use of level assessments would not allow all schools and

> student to be held to the same high standards required by the *NCLB* ACT.

According to this report, Oregon is the only state that has been approved by the U.S. Department of Education to use CAT to meet NCLB requirements. As a result of this approval, Oregon must ensure comparability of results for fixed form and adaptive test in alignment with state grade level standards, content, quality, difficulty, and subgroups. They also have to ensure that the meaning and analysis of results is consistent (Harmon, 2010).

As recent as 2010, out-of- level testing was still a point of contention. According to a consortium of test developers test used to meet the requirements of NCLB had difficulty "assessing the skills and knowledge" (Lazer et al., 2010, p. 5) of above and below level students. Many state-wide assessments accurately assess student proficient around the cut-score and above or below the cut-score. The use of only grade level content for *NCLB* requirements restricts the accurate assessments of students who score below and above the level of proficiency. The consortiums response to accurately measuring students' skills above and below proficiency levels was the use of adaptive testing. A test that is 'tailored' for individual students' ability is the most accurate way to measure students when using the new common core standards. The administration of adaptive test will allow core standards to be vertically aligned allowing for use of a growth model instead of status model to measure student performance (Lazer et al., 2010).

Although vertically aligned adaptive test can measure student growth from year to year, adaptive test can also be developed to measure on grade level content. This may be a difficult feat for test developers to undertake because of the need to develop an

extensive item pool. Other concerns, in reference to adaptive are test item complexity and innovation, the development of scoring algorithms, and scoring constructed response items (Lazer et al., 2010).

In 1980, 10 years after Frederic Lord published *Some Test Theory for Tailored Testing*, he again posited, "that in the not too distant future many mental tests will be administered and scored by the computer. . . computerized instruction will be common, and it will be convenient to use computers to administer achievement tests" (p. 150). Lord's ardent support of computer-based tests stemmed from his knowledge of how computers are able to administer multiple forms of tests to many examinees, the capability of examinees to respond to test items at different rates, and the ability for the computer to use pre-calibrated items designed specifically for the examinee (Lord, 1980). Lord further concludes that the computer was capable of administering tests items that were neither too difficult nor too easy for an examinee through the estimation of an examinee's ability level after each response. Based upon the empirical evidence of Lord as well as many others, computer adaptive tests have arrived in the 21st century (Lord, 1980).

In order to determine if CAT is feasible for a testing program, an organization should consider the following: item bank development, psychometric expertise, cost, and expected outcomes. A prudent approach for organizations considering CAT, is to conduct a simulation, such as a Monte Carlo Simulation. A Monte Carlo simulation simulates examinees to determine the feasibility of CAT. Item responses matrices from the computer program are generated to estimate theta levels and item parameters (Thompson & Weiss, 2011).

Other simulations conducted to determine the appropriateness of CAT are post-hoc and hybrid. Post-hoc simulations utilize real data to estimate examinee responses for a live CAT. A major concern to using post-hoc simulations is the issue of missing item responses. Hybrid simulations, however, addressed this problem by providing missing data from a post-hoc simulation with simulated data. The simulated data were obtained through the use of a Monte-Carlo simulation (Thompson & Weiss, 2011).

There are a series of events that occur in a live CAT administration. Each event is pertinent for the successful administration of CAT from start to finish. Flowcharts of these events are shown in Figure 3 (Hambleton & Swaminathan, 1991). The first event requires the administration of a test item of to an examinee. The examinee's response determines the level of the next test item. If the examinee responds incorrectly to a test item, an easier question would be administered while a correct response would administer a more difficult item (Hambleton & Swaminathan, 1991).

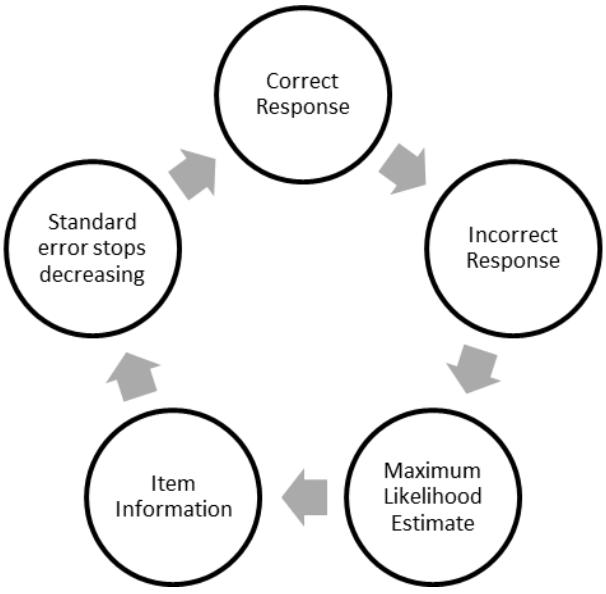


Figure 3. Computer Adaptive Test Process

Once a correct and incorrect response vector is obtained, theta was estimated using MLE. The test information for the items is gathered at the estimated ability level as well as the standard error of measurement. Then the item information is calculated for the un-administered test items at the estimated ability level. The item that would provide the most item information for the examinee would be administered next. A new ability estimate would be obtained based upon the examinee's response, thus allowing for the process to repeat itself (Hambleton, et al., 1991).

It has been an arduous and often time contentious journey to the development of computer-based assessments. Currently, policymakers and test developers have found a non-partisan way to merge conflicting ideology on high-stakes computer-based testing through the development of two consortia—Smarter Balance Assessment Consortia (SBAC) and Partnership for Readiness in College and Career (PARCC). Both consortia will deliver the next generation of assessments via computer, however SBAC states will administer CATs and PARCC states will administer linear-CBTs (US Department of Education, 2010). Testing companies are informing customers on the benefits of computer-adaptive testing (CTB-McGraw Hill, 2013). The distinguishing features of Variable-length (VL) CATs are the variability it offers in termination criteria. VL-CATs have available six different termination criteria from which to choose. These six-termination criteria range from fixed standard error of theta estimates to a change or increase in SEM to a change in theta estimates, minimum item information, and classification termination (Weiss & Guyer, 2010).

Now that CAT has been mainstreamed into the conversation of high-stakes testing, it is now time to focus our attention to the different forms of CATs. CATs are a more

precise and efficient way to measure examinee performance, however, VL-CATs offers even more. Organizations that choose CATs have the option of choosing termination points that fit the purpose of tests (Babock & Weiss, 2009). For example, CCTs classify examinees into distinct categories—Pass or Fail. If the purpose of the test is to determine whether or not an examinee meets a specified cut-point, then CCTs are the test to use due to the level of precision it can classify examinees (Parshall, et al., 2002).

CCTs along with other forms of VL-CATs have added a new twist to the debate of CATs. VL-CAT offers different points of termination for each examinee, thus bringing the issue of test precision to the forefront. Is test length a formidable factor when addressing issues of test efficiency and precision? VL-CAT provides precise measurement of an examinees ability level by specifying the termination point of the test, allowing "Well-targeted examinees to receive shorter tests than poorly targeted examinees" (Parshall, et al., p.129).

Opponents of VL-CATs argue that shorter tests are less precise. Essential components of CATs are the estimation of ability levels to determine the next item administered to the examinee. If the method used to estimate ability levels does not accurately estimate an examinees ability level then measurement precision is compromised. Cited issues with VL-CATs are the association of estimation of ability levels with tests length. Although VL-CATs are terminates at a precise level, ability levels are underestimated for students with true ability levels that are low. Thus, compromising test precision ((Parshall, et al., 2002). Babcock and Weiss (2009) found VL-CATs to estimate examinee ability level well if the standard error termination point were set at a level of 0.315 or smaller.

A fixed SE of theta estimate allows the test developer to set the SEM at a certain termination value, whereas an unfixed SEM would cause a CAT to terminate as the SEM decreases to a leveling off point. Another termination criterion is for stability of theta estimates. In rare instances, examinee responses may have to terminate due to the wrong IRT model. In this case, an increase in the SEM of theta would result in termination. The use of minimum item information would be used when the item information drops below a pre-set value.

Computerized Classification Test

A unique alternative to the other termination criteria was the classification termination. This termination utilized the mastery/classification system to terminate the test. The use of computerized classification terminates the test when the examinee's theta estimate plus or minus the confidence interval is above or below the cut-score (Weiss & Guyer, 2010). CCTs classify an examinee's score into the following categories: pass/fail, mastery/non-mastery. CCTs are used in certifying or licensing organizations. Benefits of CCTs are that only 10 items or fewer are necessary before an examinee can be classified in the pass/fail or mastery/non-mastery category. CCT items are scored using ability estimates and standard error of ability estimates. A passing score on a CCT is denoted when the examinee score is above a pre-set confidence band (Parshall, et al., 2002).

Computer Based Test vs. Paper and Pencil Test

The cost effectiveness of computer-based testing as compared to paper and pencil is debatable. School districts with a limited number of computers, find the initial cost of starting a computer-based testing program costly. School districts with an established number of computers find computer-based testing to be cheaper than PPT due to the

lower cost of scoring. For example, Indiana implemented an online end of course test in English which cost a fourth of the cost to score than paper and pencil test (Olson, 2003). Policymakers argued that CAT could not be used for accountability testing because it does not test students on grade level. Psychometricians, however, disagree with that assertion because it is the difficulty level of the test items that changes in an adaptive test and not the content (Trotter, 2003).

In 2010, Grunwald Associates conducted interviews with state and national leaders in the fields of technology and assessment to explore their beliefs, observations, and practices as it related to online assessments. The research design consisted of 81 semi-structured phone interviews from key stakeholders in technology and assessment from states where the use of online assessments is wide-spread. The interviews found that the majority of the states were implementing some form of online assessment and the majority agreed that online assessments were the wave of the future.

Test developers contend that adaptive tests align with the United State Education Department's (USED) three characteristics of summative assessments: measurement of student achievement of standards, student growth as measured by student achievement, and tracking of student growth. The USED also claims that a common core assessment system needs to provide rapid results, use technology, be able to reach a large population of students, and is able to accommodate students with disabilities. Adaptive tests meet all of these requirements; however, not all stakeholders brought into the idea of a computer adaptive assessment system.

States that allow LEA's the option of PPT or CBT further compound testing issues for policymakers. These issues are of test comparability. In order for tests to be

compared, they have to be placed on the same scale. CBT and PPT, add to this conundrum because not only are the tests on different scales, but the tests are administered differently (Olson, 2003). Comparability of CBT to PPT becomes a minor issue when the topic of 'adaptive' testing enters the discussion. Currently, CAT has not been a part of the discussion for state testing programs because federal officials state, "adaptive tests are not 'grade level test' a requirement of law" (Trotter, 2003, p.17). Policymakers contend that the use of CAT could result in lowering expectations of students who are below grade level. As a result of federal law, states such as South Dakota and Oregon had to dismantle the use of CAT as an accountability measure for NCLB (Trotter, 2003).

In November of 2007, an article in the Washington Post urged policymakers to offer more flexibility to states who wanted to implement CAT. The Washington Post cites sources supporting claims of computer-adaptive testing as an effective and accurate way to measure student ability. This claim stemmed from State representatives from Wisconsin and Oregon who proposed a bill that would allow states complete flexibility as it relates to choosing the mode of test administration (*NCLB*, 2007).

National Policymakers' 2003 stance on CAT was somewhat perplexing, since during the 1970's and 1980's the federal government began research in the area of CAT. The federal government even created grants for CAT initiatives in the area of foreign language. Policymakers and test developers are not in agreement about CAT testing on grade level. In other words, some testing experts believe that the theoretical aspects of testing do not translate into the reality of testing. Test developers do not agree with the use of CAT for accountability purposes (Trotter, 2003).

In 2001, the Oregon DOE surveyed 740 3rd graders and 730 high school students about their experiences using the Technology Enhanced Student Assessment (TESA). The survey found 3rd grade students to be more optimistic about using the computerized test in the areas of reading and math. Sixty-two percent of third graders reported the reading test was easier on the computer as opposed to 58 percent finding math easier. High school students, however, were not as optimistic. At least twenty percent of high school students said they had done better on the paper and pencil test in the areas of reading and math. Only thirty-seven percent of high school students found reading to be easier on the computer with thirty-eight percent finding math easier to use (Park, 2003).

Summary

Researchers, policymakers, test developers, and educational leaders have developed valid arguments for and against computer based testing. States have the option to deliver high-stakes tests by paper and pencil or computer, however several states have decided not to administer the computer-based assessment. States cited different reasons for their decision, but one primary reason was money (Washington, 2013). Simulation studies combined with school districts' feedback of the results is an essential component in determining the feasibility of implementing high-stakes testing on the computer.

In 2011, Thompson and Weiss proposed a framework for making CAT's operational. This framework provided five necessary steps to make CAT operational. Although their framework was constructed for use with CAT, there are many stages of the framework that can be generalized and applied to computer-based testing. An important aspect to the framework is its emphasis on "feasibility, applicability, and

planning studies" (Thompson & Weiss, 2011, p.1) prior to administering a CAT program. An overarching question for school districts is, "How feasible is it to implement a high stakes computer based testing program?"

The concept of feasibility applies to school districts having the necessary resources to implement the program. In speaking of resources, it is in regards, to personnel, computers, and time. The feasibility of implementing a computer-based testing program may be a district by district decision. Some school districts may be able to administer tests to multiple students on the computer, whereas in other districts this may not be feasible. This brings forth the question, "Is it feasible to administer tests on the computer if there is a lack of resources?" The purposed study will address the aforementioned questions by using simulated results of computer adaptive and computer classification tests. The efficiency and precision of the simulated results will be compared to the paper and pencil assessment and discussed with school test coordinators in implementing a high-stakes computer based testing program.

CHAPTER 3

METHODOLOGY

This chapter describes the research design and procedures used to conduct the study. Each section of this chapter gives a detailed description of the theoretical framework, research design, population and sample, sampling procedures, instrument descriptions, and data collection procedures. The purpose of the organization of the chapter was to provide an outline for researchers to understand test coordinators perceptions of a high-stakes computer-based testing program and the impact of a CAT simulation on their perceptions, by succinctly following the steps outlined in this chapter.

Theoretical Framework

> All knowledge, and therefore all meaningful reality as such, is contingent upon human practices, being constructed in and out of interaction between human beings and their world, and developed and transmitted within an essential social context.
>
> (Crotty, 1998, p. 42)

Crotty's definition of constructionism provided explanation for the myriad views and perceptions associated with the use of computer-based tests. The perceptions of the aforementioned were constructed by their own experiences. In some instances, views of computer-based tests were constructed due to lack of experiences. In order to understand the lack of computer-based tests in 21st century testing, one must construct meaning of this phenomenon not from solely an objective or subjective viewpoint. Instead, the phenomenon must be observed relevant to an individual's experiences (Crotty, 1998).

Computer-based testing devoid of any human interaction would be completely objective. "In assessment, performance is not 'objective'; rather, it is construed

according to the perspectives and values of the assessor, whether the assessor is the one who designs the assessment and its 'objective' marking scheme or the one who grades open-ended performances (Gipp, p.370)." It was the norm to view psychometrics, the study of test and test theory, as objective, positive, and experimental (Crotty, 1998). Whereas, Broadfoot (1994) argued that assessments are "not an exact science" (cited from Gipp p. 370). A case study design was employed to collect qualitative data because it provided an in depth investigation of the phenomenon through real experiences (Yin, 2009).

To explain the phenomenon of large scale testing to an audience of parents is a feat not even the most well versed testing expert can claim. The testing phenomenon is just that, a phenomenon---something you inevitably have to experience for yourself. Many individuals have had some experience with testing, albeit classroom, diagnostic, or standardized. Regardless of the tests, their perceptions of the tests are based upon their personal experiences.

"Perception is a negotiation among patterns we detect in the environment and patterns of accumulated experience" (Mislevy, p. 273); the essence of phenomenology is the understanding of the phenomenon from the perspective of the individual experiencing the phenomenon. In the case of testing, individuals have acquired preconceived notions based upon their experiences or the experiences of others. This is extremely important as it relates to understanding individuals perception of implementing computer adaptive testing from a phenomenological stance. When viewed through the lens of phenomenology, one has to 'bracket' (Moustakas, p. 97) their present understanding of the phenomenon to create new meaning (Crotty, 1998; Moustakas, 1994).

A phenomenological framework allows for the researcher as well as the participants to lay aside all preconceptions of testing to introduce a new testing phenomenon --- CATs. Exploring the phenomenon of CATs through the use of simulations and perceptions will constitute a transformation in the way test are administered.

Research Design

> Case studies are a common way to do qualitative inquiry. Case study research is neither new nor essentially qualitative. Case study is not a methodological choice but a choice of what to be studied. (Stake, 2005, p. 443).

This case study investigated the perceptions of school test coordinators on a computer adaptive testing program. A multiple case studies design was used to collect and analyze data. In addition to the design consisting of multiple cases, embedded within each case was a method of data collection and analysis, thus classifying the design as a multiple-case embedded design (Yin, 2009).

This focus of this case study was a school district and a computer-based testing program. The study examined computer-based testing across several schools, in what Yin (2009) described as a multiple case study. Replication logic was used to explain the use of a multiple case study design. The rationale for use of a multiple case study is similar to conducting multiple experiments. Each time an experiment is repeated, and if similar or exact results are obtained, the hypothesis, which is tested, will be substantiated. As an embedded design, this case study followed the logic of replication. The same research procedures were replicated in each school to address the research questions (Yin, 2009).

Triangulation of data sources is an important part of case study research. The interweaving of multiple data sources alleviates issues with construct validity

since all sources are examining the same phenomenon (Yin, 2009). The primary method of data collection was five open-ended interviews conducted with school test coordinators at each participating school. However, a variety of data sources were used to validate data obtained from the interviews. Interviewees were asked to submit artifacts of computer testing procedures and tests administration schedules. Another data source included in the study was an observation of computer resources. The CATSim Program (Appendix D) was used as elicitation material within the interview. The results from CATSim were used to elicit a response from the interviewee in regards to their perception of implementing a high-stakes computer-adaptive testing program.

The interviewees received 15 open-ended interview questions prior to the day of the pre-scheduled interview. Each interviewee was informed that the interview would be digitally recorded and transcribed verbatim. The interviewee had an opportunity to review all contents of the transcription. Interviewees were allowed to correct any errors in transcription.

Data were collected through interviews and documents. The following procedures were followed for analysis (1) data were read and transcribed while notes were written in the margins; (2) data were categorized, colored coded, and organized into categories to make meaning; (3) similar categories were connected to make themes; (4) a narrative was written to summarize the data; (5) meaning was constructed from the narrative (Creswell, 2003). For the purpose of this multiple case study, an embedded analysis was conducted for each school using the previously discussed procedures. The

themes for each school were then collectively analyzed for similarities and/or differences (Yin, 2009).

Population and Sampling

The data for this study were collected from five out of thirty elementary schools located within an urban school district located within the continental United States. The sampling techniques used within the study were derived from both qualitative and quantitative approaches. Specifically, the number of examinees chosen for the CAT simulation was based upon the minimum number of item responses required to optimize simulation results. Thus, a total of 586 examinees' item responses were collected from five elementary schools for the study. The five elementary schools were selected based upon the following criteria: (1) identified as a Kindergarten through five school and (2) the reported number of fifth graders administered the Social Studies Benchmark Assessment greater than or equal to 100. A purposeful sampling technique was used to choose interviewees following the selection of the five elementary schools. The test coordinators from each of the identified schools were selected to participate in the study due to their knowledge, experience, and expertise in the field of elementary school testing.

This research study utilized a quantitative technique, CAT simulation, to elicit test coordinators' responses on implementing a computer-based testing program. The simulation was chosen to demonstrate the efficiency and precision of a CAT if used as part of the District's testing program. Simulation results from the CATSim Program were analyzed to determine the efficiency and precision of the CAT and CCT.

The qualitative and quantitative data collected in this study were triangulated for evidence and support of the research questions and purpose of the study. The triangulated data included results from the simulation, interviews, documents, and photos provided as supporting evidence of interview responses. Using real item responses, a post-hoc simulation was conducted to explore CAT efficiency and precision using different termination criteria. Test length was observed to address CAT efficiency and estimated CAT theta and true theta values were compared to address CAT accuracy and precision.

Table 1

Participant Demographics

Participants	Gender	Age	Race/Ethnicity
A	Male	45-54	African American
B	Female	35-44	African American
C	Female	35-44	African American
D	Female	45-54	African American
E	Male	35-44	African American

Table 2

Participant Background

Participants	Certification	Highest Degree	Years of Experience
A	Administration	Masters	11 to 20
B	Administration	Specialist	0 to 5
C	Administration	Masters	0 to 5
D	Administration	Doctorate	6 to 10
E	Administration	Doctorate	0 to 5

Tables 1 and 2 contain demographic, work experience, and degree information from the five interview participants. All interview participants are African-Americans with 0 to 20 years of experience, 2 male and 3 female, who hold a certificate in Educational Administration, and have advanced degrees. Data used for the CATSim Program were obtained from fifth grade examinees' Social Studies Benchmark Assessments administered during the month of October. Item responses from a population sample of 586 examinees were obtained from the paper and pencil administration of the Social Studies Benchmark. Schools were purposefully chosen to participate in the study based upon the following criteria: 5th grade class size equaled 100 students plus or minus five students and the school administrator serves as the test coordinator. The sample size for examinees of 500 was chosen because the minimum number of item responses needed to run the simulation program was 500. Interviews were conducted with school Test Coordinators from each of the five elementary schools where student responses were obtained for the simulation.

Instrumentation

The CATSim Program was used to demonstrate the efficiency and precision of CAT. This program utilized examinees' item responses from a paper and pencil administration of a Social Studies Benchmark Assessment. Examinees' responses were then used to simulate a CAT.

1. Item responses were obtained from a 30 item Social Studies Assessment administered to 586 examinees.

2. XCalibre 4 (Appendix D) was used to estimate item parameters. The program removed six items from the 30 items due to low point bi-serial numbers.
3. Item parameters estimated from XCalibre 4 were then uploaded into the CATSim Program.
4. MLE was used to estimate the theta level for each examinee.
5. Maximum Information function was used to select the next test item administered to the examinees.
6. Two VL-CAT simulations were run using the following termination points: standard error of measure and computerized classification.

Examinees were administered the paper and pencil version of a Social Studies Benchmark Assessment. This assessment was developed from a team of content experts within the district. A test blueprint was created to guide the development of the assessment as well as the selection of items used from the item bank. A paper and pencil version of the Social Studies Benchmark Assessment was administered to all fifth grade students within the school district. Each examinee was administered a 30 item social studies test and with a time constraint of sixty minutes. The sixty-minute time constraint was consistent with time constraints implemented during statewide assessments. Examinees recorded their responses on an answer document. Each examinee's answer document was scanned and uploaded into a data management system. Item parameters were then estimated using the XCalibre 4 Program. The total number of items decreased from 30 to 24 due to low point-bi-serial numbers.

Procedures

A simulation program was chosen to measure test efficiency and precision due to the usefulness of simulation programs in predicting the outcomes of proposed CAT design (Davey and Pitoniak, 2006). The CATSim Program was used to measure the precision of CAT by comparing the estimated CAT theta values from each simulation with the true thetas values obtained from the 586 examinees. Examinees with true theta values of -4 to 4 were examined for the study.

Simulation results were compared using standard error and classification termination points. The first VL-CAT simulation was set to terminate when "the change in successive standard errors was less than or equal to 0.005" (Weiss & Guyer, p. 29). The classification test terminated when the theta estimates fell above or below the confidence interval. The confidence interval was set plus or minus 2.00 standard errors of the cutoff value of 1.000 (Weiss & Guyer, p. 29). Decision accuracy for classification termination increases with easier test items (Luecht, 2006). Simulated results were produced from the CATSIM program using the listed procedures (Weiss & Guyer, 2010):

1. All examinees started with an initial theta level of 0.
2. Maximum likelihood estimates were used to obtain theta estimates. Theta estimates were only obtained when the examinee answered 1 item correct and 1 item incorrect. To increase an examinee's response pattern of at least 1 correct and 1 incorrect, a step size of 0.5 was selected. The purpose of the step size was to ensure that the examinee obtains a response pattern of 1 correct and 1 incorrect by increasing the difficulty level of the next selected

test item.

3. Maximum Fisher information was used to determine the next test to ensure the amount of information that was provided by the item. Administered items were re-entered into the item pool.
4. Two CAT simulations were conducted using different termination points. First, the "Variable Termination" tab was selected for VL-CAT, and then the termination criteria, "terminate when the change in successive standard errors is less than or equal to .005" (Weiss & Guyer, 2010, p. 29). The same procedures were then followed for CCT by selecting the termination criteria of "terminate when the theta estimate plus or minus 2.00 standard errors is above or below a theta cutoff value of 1.00" (Weiss & Guyer, 2010, p. 29).

The results from the simulation were then used during the five, 60 minute interviews conducted with the Test Coordinators from each school. Interview questions were constructed to answer the research questions by having interviewees describe their experiences with testing. Appendix A lists the interview questions. Individual interviews were "guided conversations rather than structured queries" (Yin, p. 106). Questions were constructed to guide examinees responses on their perceptions of high-stakes computer-based tests (Yin, 2009).

Research participants were asked to provide any supporting evidence for their responses. Appendix A documents computer-based testing schedules implemented by test coordinators for District Benchmark Assessments. Observational evidence was used to provide a better understanding of the resources necessary to implement a testing program (Yin, 2009). Evidence was collected

through photography, to support interviewee responses regarding computer resources.

CHAPTER 4

RESULTS

This chapter provides a detailed account of the data collected for the research study. The focal point of Chapter 4 was the analysis of various data sources. In order to understand the connection of each data source to the case study and more importantly the proposed research questions the analysis begins with the CATSim Program, followed by interviews, then photographs, and ending with documents. The detailed description of the analysis will allow the reader to understand the triangulation of the data sources.

The purpose of the research study was to address the questions: (1) Are the efficiency and precision of computer-adaptive tests and computer classification tests paper equivalent or superior to those of paper and pencil benchmark tests? (2) What are test coordinators' perceptions of administering high-stakes tests on the computer? (3) To what extent, if any, did computer simulation results elicit a change in test coordinators' perceptions of administering high-stakes tests on the computer? Each analysis consists of a description of the data source, an analysis of the data source, and summary of the analysis. Each analysis was explained in terms of the school/case in which the data were collected. The exception to this was the interviews, where the analysis was arranged in terms of emerging themes. Tables, figures, and documents used in the analysis are found in the appendices or throughout the chapter to add clarity to the analysis.

Results from the CATSim Program were used as documentation to show school test coordinators the efficiency and proficiency of a CAT. The simulation used combined benchmark data obtained from 586 examinees attending schools of the interview participants. Table 1 was used to discuss the efficiency and proficiency of a VL-CAT

and a CCT with school test coordinators interview during the interview. Table 1 results provide details of both variable-length simulation as well as computerized classification simulation. The VL-CAT simulation was run using a variable-length termination point where the change in successive standard errors was less than or equal to .005; whereas a CCT termination point was set at a theta level of ±2.00 standard errors above or below the cutoff theta level of 1.00. The histogram displayed in Figure 1 shows a graphic representation of the frequency of items administered to examinees for a VL-CAT and a CCT. Table 2 lists descriptive statistics for a VL-CAT and a CCT based upon the number of items administered to each participant (N). In Table 3, a statistical summary is provided of the ability estimates for the full bank of items, a VL-CAT, and a CCT. The standard errors of theta estimates for a VL-CAT and CCT as compared to the full item bank are listed in Table 4. A comparison of the difficulty levels of items for both VL-CAT and CCT is listed in Table 5. Table 6 lists the response vectors, estimated CAT theta, and SEM for two hypothetical examinees, *Examinee A* and *Examinee B*. Item information and standard error for each theta level is located in Appendix C.

Simulation Results

Results from the CATSim Program are displayed in Tables 3-8 and Figure 5. The data results displayed in Tables 3-8 were based upon the following: 586 examinees, a 24 item exam, and VL-CAT and CCT termination criteria. The number of items administered to examinees was different for each simulation. For example, the VL-CAT stopped administering items when the change in successive standard errors was less than or equal to .005. The CCT simulation stopped administering items when the estimated theta level ±2.00 the confidence interval fell above or below the cutoff theta level of 1.00.

Table 3 summarizes the number of items administered and the frequency of distribution for the VL-CAT and CCT.

Table 3

Number of Items Administered to Examinees

	VL-CAT			CCT		
Item	N	Cum. N	Percent	N	Cum. N	Percent
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	47	47	8.02	206	206	35.1
6	34	81	5.80	9	215	1.54
7	0	0	0	18	233	3.07
8	0	0	0	10	243	1.71
9	31	112	5.29	19	262	3.24
10	1	113	0.17	10	272	1.71
11	25	138	4.27	11	283	1.88
12	6	144	1.02	11	294	1.88
13	32	176	5.46	13	307	2.22
14	16	192	2.73	17	324	2.90
15	30	222	5.12	4	328	0.68
16	38	260	6.49	6	334	1.02
17	22	282	3.75	17	351	2.90
18	34	316	5.8	7	358	1.20
19	42	358	7.17	1	359	0.17
20	55	413	9.39	2	361	0.34
21	21	434	3.58	1	362	0.17
22	73	507	12.5	5	367	0.85
23	44	551	7.51	1	368	0.17
24	35	586	5.97	218	586	37.2

VL-CAT Simulation

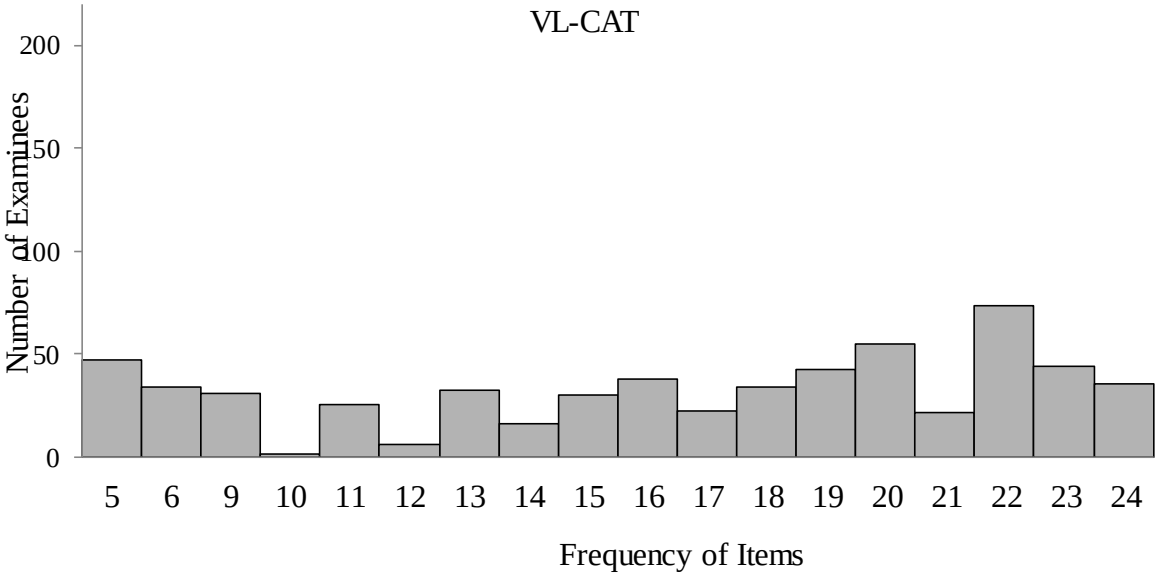
Out of a 24 item test, the minimum number administered to a student before the test would terminate was five; whereas the maximum number of items administered to students prior to termination was 24. There were 47 students administered the minimum number of items prior to the test terminating, and 35 students were administered 24 items prior to the test terminating. The percentage of students who received five or 24 items was 8% and 6%, respectively. The number of items administered to the highest percentage (12.5 %) of students was 22. Only one student was administered 10 items which constituted less than 1% of the population. Table 4 is a summary of the descriptive statistics for the number of items administered using a termination point of .005 for VL-CAT and plus or minus 2.00 standard errors above or below a theta cutoff level of 1.00. The selected theta cutoff-level of 1.00 was chosen because of the score ranges and simulation results from the benchmark assessment conducted prior to the study.

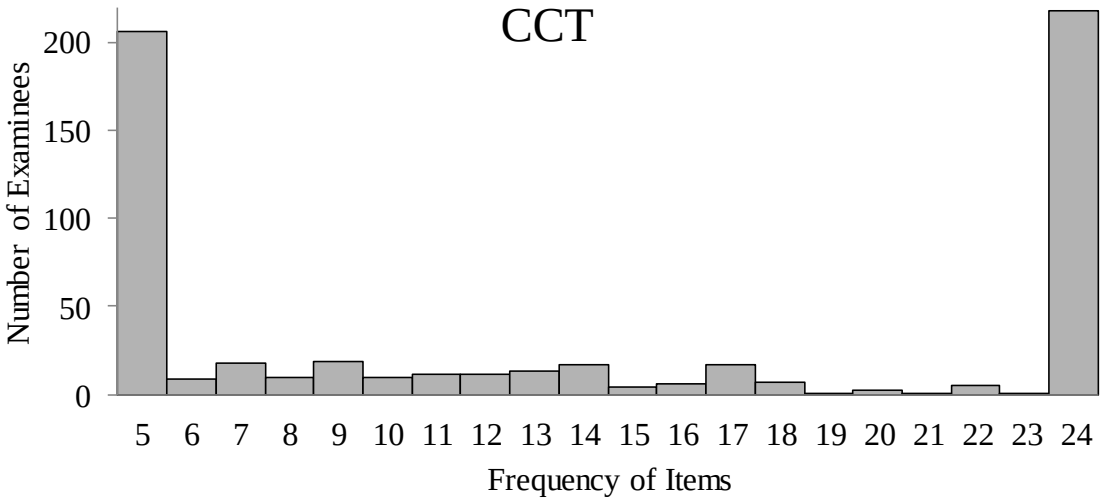
CCT Simulation

For the CCT simulation, a minimum of five items were administered to all examinees and with a maximum of 24 items. There were a total of 206 examinees who received the minimum number of items (5) and a total of 218 examinees who received the maximum number of items (24). The percentage of examinees administered the minimum and maximum number of items was 35.15% and 37.20%, respectively. Less than 30 percent of the number of items administered to examinees ranged from 6 to 23 items. There were three examinees who were administered either 19, 21, or 23 items, which constituted less than the 1 percent of the total number of items administered.

Figure 4

Frequency of Administered Items





Summary of Frequency of Administered Items

Figure 4 displays the distribution of administered items across examinees. The frequency of items administered was more evenly distributed across examinees for VL-CAT than CCT. The frequency of the items administered from 5-24 was evenly distributed among the examinee population of 586 in VL-CAT. Item frequency for CCT differed drastically from VL-CAT with approximately half of the examinees administered five items and the other receiving all 24 items.

Table 4

Summary of Descriptive Statistics for the Number of Items Administered

	Mean	SD	Variance	Minimum	Maximum	Range
VL-CAT	15.28	5.72	32.68	5.00	24.00	19.00
CCT	14.00	5.63	31.67	5.00	24.00	19.00

Table 5

Summary Statistics of Theta Estimates

	Theta Estimates		
	Full Item Bank	VL-CAT	CCT
Mean	0.61	0.49	0.33
Standard Deviation	0.58	0.98	0.81
Minimum	-4.00	-2.50	-1.14
Maximum	2.16	2.50	2.37
Root Mean Square Difference		0.65	0.46
Correlation		0.78	0.84

VL-CAT Simulation

A summary of theta estimates is shown in Table 5 for the full item bank, VL-CAT, and CCT. Theta estimates for the full item bank are estimated by administering examinees the full item bank. Theta estimates are calculated using a full item bank every time an examinee is administered an item. The next item is then administered to an examinee based upon the amount of information given at the estimated theta. This iterative process continues until the full bank of items is administered. This process is the same for VL-CAT and CCT. As a result, the mean theta estimate for the full item bank (0.61) was found to be slightly greater than VL-CAT (0.485). However, the standard deviation of theta estimates was greater for VL-CAT (0.98) greater than the full item bank (0.58). The minimum and maximum theta estimates were found as -4.00 and 2.16, respectively, whereas VL-CAT's minimum theta was -2.50 with a maximum of 2.50.

CCT Simulation

Table 5 also shows that the mean theta estimate (0.61) was greater than the CCT theta (0.33). The standard deviation of the standard deviation of the theta estimate for the full item bank (0.58) was less than the CTT theta (0.81). The minimum CCT theta was -1.14 and maximum 2.37 while the minimum theta estimate for the full length test was -4.00 and the maximum theta estimate was 2.16. A strong, positive correlation exists between theta estimates for full bank CAT thetas and CCT thetas (r = 0.84). The minimum theta (-1.14) for CCT was higher than the minimum theta (-2.50) for VL-CAT. However, the maximum theta (2.50) VL-CAT was higher than the CCT maximum theta (2.37). The mean theta (0.49) for VL-CAT was higher than the mean theta (0.33) for CCT due to higher overall theta values of VL-CAT.

Table 6

Standard Errors of Theta Estimates

	Theta Estimates		
	Full Bank Theta	VL-CAT Theta	CCT Theta
Mean	0.278	0.563	0.430
Standard Deviation	0.117	0.729	0.189
Minimum	0.251	0.251	0.251
Maximum	3.000	3.000	3.000
Correlation		0.198	0.612

VL-CAT Simulation

Table 6 shows that the average standard error for full bank theta and VL-CAT theta was 0.278 and 0.563 respectively. The higher standard error for VL-CAT theta denotes the VL-CAT had less test precision. The larger standard of error is due to the shortened length of the test with the average number of test items administered for VL-CAT at 15.28. Although VL-CAT was a shorter test, the number of items administered for this test was evenly distributed. The number of examinees administered 24 items was less than the number of examinees administered 15 or fewer.

CCT Simulation

In the Table 6 comparison of the mean standard error of full bank theta to CCT theta, the mean is higher for CCT theta. The larger standard error of measure was expected for CCT because of the shorter length of the test. Two hundred and six (206) examinees test terminated after being administered five items. There was a strong,

positive correlation (0.612) between full bank and CCT standard error of measures. So, as the standard error for the full item bank increases the CCT error values also increase.

Table 7

Frequency of Item Difficulty

	VL-CAT		CCT	
Item	b	Frequency	b	Frequency
1	0.00	586	0.00	586
2	0.81	411	0.81	274
3	0.59	433	0.59	296
4	0.98	313	0.98	276
5	0.89	427	0.89	283
6	0.94	364	0.94	275
7	0.00	474	0.00	547
8	1.12	287	1.12	277
9	0.71	446	0.71	298
10	0.80	422	0.80	286
11	0.75	455	0.75	291
12	1.17	218	1.17	272
13	0.00	445	0.00	539
14	1.06	279	1.06	269
15	0.50	513	0.50	400
16	0.58	417	0.58	308
17	0.62	420	0.62	292
18	0.97	307	0.97	267
19	0.29	427	0.29	499
20	0.96	339	0.96	268
21	0.90	385	0.90	278
22	0.45	470	0.45	400
23	0.92	352	0.92	267
24	0.00	366	0.00	489

VL-CAT Simulation

Table 7 displays the frequency of item difficulty. The initial theta level for all examinees was 0.000. As a result, each examinee was administered item number 1, with a difficulty level of 0.000. Subsequently, item numbers 7, 13, and 24 all had difficulty levels of 0.000 with item number 7 administered most frequently and items 13 and 24 following closely with 445 and 366 items administered, respectively. Item 12 was administered the least frequently, but had the highest difficulty level at 1.169. The lower the level of difficulty the higher the frequency rate in which an item was administered. Items that had a difficulty level of one or close to 1 had the lowest frequency numbers. The highest level of item difficulty was 1.17. The frequency of administration for this test item was also the lowest, which indicates examinees were more likely to be given easier items from the item bank.

CCT Simulation

In Table 7, the frequency at which an item was administered was determined by the level of theta. In CCT as well, items with low difficulty levels were administered more often. For example, 0.000 to 0.496 were the most frequently administered items. The frequency of the item administration ranged from 400 to 586. However, the more difficult items, with a 1.123 and 1.169 level of difficulty had among the lowest administrations, 268 and 269, respectively. The negative correlation between item difficulty and frequency in item administration was also evident in CTT. The VL-CAT and CCT were similar in the number of items that were administered at lower and higher frequency levels. For example, at a difficulty level of 1.169, the number of items administered for CCT was 272, whereas VL-CAT was 218. For items with a 0.000 level

of difficulty, both tests had an administration frequency as low as 366 to as high as 586.

Item Information

Appendix C lists the information function and standard errors for each theta level. The amount of information a test provides at a certain level of theta is inversely proportional to the standard error of measure. Thus, as the standard error of measure increases, the amount of information provided for the test question decreases. As shown in Appendix C, the lower theta values correlate with a low test information and higher standard error of measurement. As the standard error of measurement increases, the level of precision decreases. As the ability level of the examinee increased, the precision level of the test also increased. So, the test was able to provide more precise information for examinees at a higher theta level than examinees at a lower theta level. Maximum information was provided for an examinee with a theta level of .70, with a standard error of measurement at 0.2508.

Table 8

Full Response Vectors for Examinees A and B

Examinee A (theta=0.8898)				Examinee B (theta=2.1607)			
Item	Response	CAT Theta	SEM	Item	Response	CAT Theta	SEM
1	0	-0.5000	3.0000	1	1	0.5000	3.0000
7	1	0.0000	0.8319	15	1	1.0000	3.0000
13	1	0.4077	0.7204	4	1	1.5000	3.0000
22	1	0.7746	0.6875	12	0	1.3945	0.7096
11	0	0.4897	0.5553	8	1	1.6648	0.6774
15	1	0.7135	0.5242	14	1	1.8420	0.6574
9	1	0.9180	0.5059	18	1	1.9647	0.6445
21	1	1.1094	0.4946	20	0	1.5497	0.4909
8	1	1.3003	0.4887	6	1	1.6480	0.4797
12	0	1.1189	0.4263	23	1	1.7293	0.4715
14	0	0.9799	0.3892	21	1	1.7988	0.4651
4	0	0.8658	0.3641	5	1	1.8599	0.4601
5	0	0.9718	0.3522	2	1	1.9094	0.456
6	1	1.0674	0.3427	10	1	1.9542	0.4527
18	1	1.1536	0.335	11	1	1.9924	0.4499
20	1	1.0529	0.3163	9	1	2.0263	0.4476
23	0	1.1239	0.3095	17	1	2.0544	0.4457
2	1	1.1812	0.3038	13	1	2.0796	0.444
10	1	1.0862	0.2892	16	1	2.1033	0.4436
17	0	1.1289	0.2844	22	1	2.1219	0.4414
3	1	1.1667	0.2803	19	1	2.1357	0.4405
16	0	1.0752	0.2687	7	1	2.1442	0.4398
19	0	0.9827	0.2599	13	1	2.1525	0.4393
24	0	0.8898	0.2534	24	1	2.1607	0.4387

Summary of Response Vectors for Examinees A and B

Table 8 displays response vectors for two examinees with a full bank theta of 0.8898 and 2.161, respectively. The criteria used to select examinees found in Table 8, were based upon the proximity of the examinees' theta level to the cutoff theta level of 1.00. *Examinee A* (0.8898) represents the theta level closest to the cutoff theta of 1.00,

whereas *Examinee B* (2.161) represents the theta level furthest from the cutoff theta of 1.00. The table lists items administered, response vectors, CAT theta, and SEM of a full item bank. The response vectors were obtained from a CAT simulation without a specified termination point. If a termination criterion were specified using either VL-CAT or CCT, the examinees in Table 6 would not have been administered all items displayed. Applying termination points to the data listed in Table 8 would result in each examinee's test ending with a different number of items. Tables 9 and 10 display theta estimates, SEM, and termination values for all items administered beginning with item number one. The VL-CAT simulation, for example, terminated after the SEM for each examinee stopped decreasing by .005; then *Examinee* A (theta =0.8898) would be administered 19 items; whereas *Examinee B* (theta=2.1607) would have 11 items. The difference in the number of items administered to examinees was because the test terminated when the change in the SEM was less than or equal to .005.

Table 9

VL-CAT Termination Criterion for Examinees A and B

Examinee A (0.8898)			Examinee B (2.1607)		
Theta Est.	SEM	Diff. SEM	Theta Est.	SEM	Diff. SEM
-0.5000	3.0000	0.0000	0.5000	3.0000	0.0000
0.0000	0.8319	0.1115	1.0000	3.0000	0.0000
0.4077	0.7204	0.0329	1.5000	3.0000	2.2904
0.7746	0.6875	0.0131	1.3945	0.7096	0.0322
1.1028	0.6744	0.0019	1.6648	0.6774	0.0200
1.4394	0.6725	0.1460	1.842	0.6574	0.0129
1.1199	0.5265	0.0637	1.9647	0.6445	0.1536
0.9219	0.4628	0.0205	1.5497	0.4909	0.0112
1.0802	0.4423	0.0151	1.6480	0.4797	0.0082
1.2176	0.4272	0.0379	1.7293	0.4715	0.0064
1.0576	0.3893	0.0251	1.7988	0.4651	0.0050
0.9340	0.3642	0.0123	1.8599	0.4601	0.0041
1.0385	0.3519	0.0098	1.9094	0.456	0.0033
1.1263	0.3421	0.0185	1.9542	0.4527	0.0028
1.0178	0.3236	0.0141	1.9924	0.4499	0.0023
0.9214	0.3095	0.0083	2.0263	0.4476	0.0019
0.9904	0.3012	0.0067	2.0544	0.4457	0.0017
1.0464	0.2945	0.0056	2.0796	0.444	0.0014
1.0926	0.2889	0.0047	2.1033	0.4426	0.0012
1.1332	0.2842	0.0114	2.1219	0.4414	0.0009
1.0406	0.2728	0.0041	2.1357	0.4405	0.0007
1.0752	0.2687	0.0088	2.1442	0.4398	0.0005
0.9827	0.2599	0.0065	2.1525	0.4393	0.0006
0.8898	0.2534	0.2534	2.1607	0.4387	0.4387

Table 10

CCT Termination Criterion for Examinees A and B

Examinee A (0.8898)				Examinee B (2.1607)			
Theta Est.	SEM	LL	UL	Theta Est.	SEM	LL	UL
-0.5000	3.0000	-6.5000	5.5000	0.5000	3.0000	-5.5000	6.5000
0.0000	0.8319	-1.6638	1.6638	1.0000	3.0000	-5.0000	7.0000
0.4077	0.7204	-1.0331	1.8485	1.5000	3.0000	-4.5000	7.5000
0.7746	0.6875	-0.6004	2.1496	1.3945	0.7096	-0.0247	2.8137
1.1028	0.6744	-0.2460	2.4516	1.6648	0.6774	0.3100	3.0196
1.4394	0.6725	0.0944	2.7844	1.842	0.6574	0.5272	3.1568
1.1199	0.5265	0.0669	2.1729	1.9647	0.6445	0.6757	3.2537
0.9219	0.4628	-0.0037	1.8475	1.5497	0.4909	0.5679	2.5315
1.0802	0.4423	0.1956	1.9648	1.648	0.4797	0.6886	2.6074
1.2176	0.4272	0.3632	2.072	1.7293	0.4715	0.7863	2.6723
1.0576	0.3893	0.279	1.8362	1.7988	0.4651	0.8686	2.729
0.934	0.3642	0.2056	1.6624	1.8599	0.4601	0.9397	2.7801
1.0385	0.3519	0.3347	1.7423	1.9094	0.456	0.9974	2.8214
1.1263	0.3421	0.4421	1.8105	1.9542	0.4527	1.0488	2.8596
1.0178	0.3236	0.3706	1.665	1.9924	0.4499	1.0926	2.8922
0.9214	0.3095	0.3024	1.5404	2.0263	0.4476	1.1311	2.9215
0.9904	0.3012	0.388	1.5928	2.0544	0.4457	1.163	2.9458
1.0464	0.2945	0.4574	1.6354	2.0796	0.4440	1.1916	2.9676
1.0926	0.2889	0.5148	1.6704	2.1033	0.4426	1.2181	2.9885
1.1332	0.2842	0.5648	1.7016	2.1219	0.4414	1.2391	3.0047
1.0406	0.2728	0.495	1.5862	2.1357	0.4405	1.2547	3.0167
1.0752	0.2687	0.5378	1.6126	2.1442	0.4398	1.2646	3.0238
0.9827	0.2599	0.4629	1.5025	2.1525	0.4393	1.2739	3.0311
0.8898	0.2534	0.3830	1.3966	2.1607	0.4387	1.2833	3.0381

CCT Simulation

A termination point using CCT could also be determined from the data in Table 8. In this instance, the test was set to terminate when the theta estimate is plus or minus 2.00 standard errors above or below the theta cutoff level of 1.00. Table 8 displays the theta estimates, SEM's, and confidence intervals for CCT using the previously mentioned termination criteria. If the CCT termination point were applied to *Examinees' A* and *B*

CAT theta estimates and SEM's then, the following number of items would be administered: *Examinee A*-24 items and *Examinee B*- 14 items. After item 14 was administered to *Examinee B,* the test terminated and *Examinee B* was classified as *passing.* This was due to the location of the confidence interval which was above the cutoff theta level of 1.00. In summary, as examinee's theta levels approached the cutoff value of 1.00, more items were administered. In contrast, as examinee's theta level moved away from the cutoff value of 1.00, fewer items were administered.

Interviews

Interviews were conducted with test coordinators from the schools where students' item responses were obtained. Each test coordinator provided a powerful and detailed account of his or her testing experiences. The following processes were used to analyze data collected from the interviews: transcription, reading and note-taking, color coding system, creating categories, and establishing themes. Interview responses were divided into three categories: participant perspectives, planning and implementation, and rules and guidelines. As a result of this process, there were three overarching themes that emerged related to testing. The themes were: (1) security, (2) emotions, and (3) management. Figure 6 shows the categories and themes that emerged from the analysis.

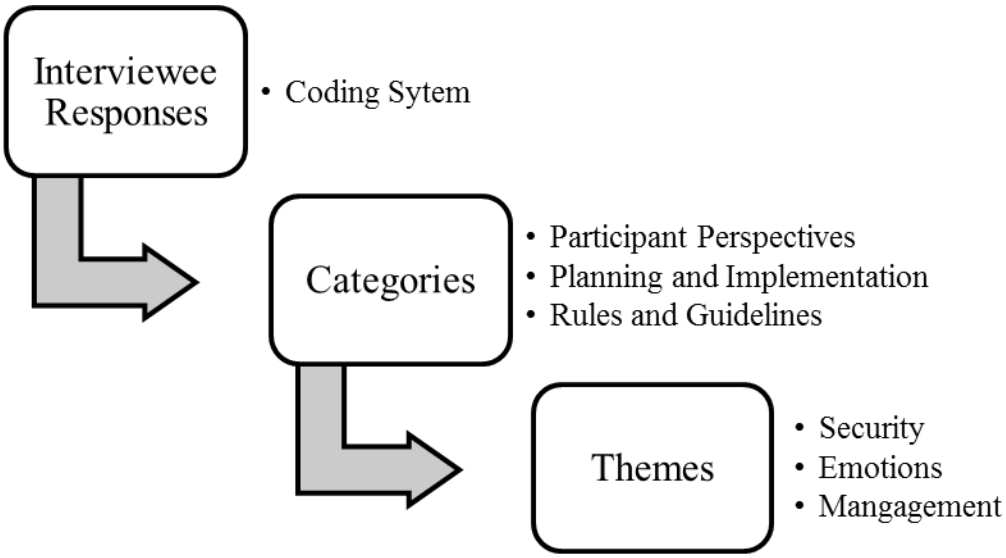


Figure 5. Organization of Analysis

Figure 5 displays how I analyzed the interview transcripts. I realized how test coordinators' experiences shaped their view of computer-based testing. It was through the process of data analysis that led to the major themes of security, emotions, and management. This process involved the initial reading of the transcripts prior to any analysis. This initial reading of the transcripts was essential to the development of the themes. Test coordinators gave a rich description of their testing experiences that allowed me to visualize the process from their point of view. Although test coordinators provided vivid recollections of their experiences, it was the emotions that evolved from these experiences that gave me a true understanding of their perceptions.

As I read each transcript, I applied a coding system which involved the highlighting of words or phrases that were the same, similar or opposite in meaning. Words were then grouped and analyzed according to the following categories: participant perspectives, planning and implementation, and rules and guidelines. To determine the themes that appeared in the study, I asked myself questions such as, (1)

What is the tone of these words? (2) What is the overall meaning of the words in the context of the study? (3) How are these words connected? (4) Is there one word that best defines the list of words within each category?

Once the themes were identified, the occurrences of the themes were associated with the test coordinators' pre-simulation experiences. Most test coordinators had limited experience, if any, with computer-based testing. Additionally, test coordinators' experience with VL-CATs and CCTs did not exist. As a result, responses elicited from the simulation were shaped by their lived experiences of high-stakes testing. Test coordinators were able to understand and form perceptions of computer-based testing by associating it to what was their current understanding of testing --- security, emotions, and management.

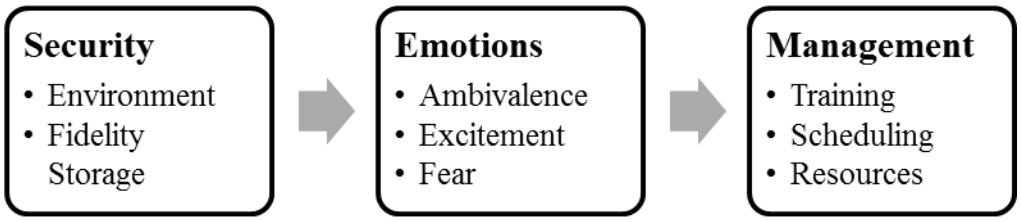


Figure 6. Major Themes

Figure 6 displays the three themes that emerged from the analysis. Security, however, was the overarching theme. Test security affected the emotions of test coordinators as well as how they managed their time, personnel, and resources. Certain factors associated with test security, such as erasures, irregularities, violations, and fidelity produced emotions of anxiety, frustration, and stress. As a result of these emotions, testing processes that involved training, scheduling, and resources were meticulously implemented. Test coordinators' perception of high-stakes testing differed from that found within the literature review. As test coordinators recounted their experiences with high-stakes testing, it was evident that testing was more than just an informal notion of assessing student learning outcomes.

Testing was a secure process, with stringent guidelines that evoked emotions similar to what was expected of the test taker and not the test coordinator. The process of coordinating high-stakes tests required a lot of organizing, pre-planning, training, and resources (personnel). Due to the high-stakes nature of the tests, ensuring test fidelity

was paramount. As a result, test security was a primary focus in the process of coordinating a high-stakes test, which was the basis for certain emotions.

The emotions of test coordinators varied as it related to different aspects of their experiences with high-stakes testing and computer-based testing. *Participant A* described the process as overwhelming because testing was only one of the duties associated with their job description, and it was very important to implement the test with fidelity. The process was also viewed as stressful and caused anxiety due to the amount of organization required and the impact of the results. Participants described their emotions below:

Participant A

> With all the other duties and responsibilities . . . it can prove to be very overwhelming at times. At times without the proper assistance it makes me feel like I am overwhelmed.

Participant B

> The first adjective that comes to mind is anxiety. It is a very busy time of year that causes stress on my family, however if I put time in at the front end . . . I know it will take away the stress of having irregularities. I take my time and plan everything out so my anxieties go away.

Participant C

> It can be extremely stressful. You have to be extremely organized. You have to be extremely knowledgeable, a lot of patience. It can be a time that is very stressful to teachers with everything going on with new procedures, policies.

Participant D

> My experience as a test coordinator can be one of stress, but what I found, is that the more organized you are, the more familiar you are with the requirements and testing the more familiar you are with the requirements and the teachers and students it makes it a lot less stressful.

Participant E

It was very a nerve racking experience in terms of getting the results back.

Test Coordinators' Perceptions of High-Stakes Testing

Participant A

> I think when we are looking at high stakes testing, I think one of the things that I am concerned with is test security. Even though, we had locks changed where tests are kept secured, it is still not like a lot of other schools where they have the ability to put them in their safe where they can safe guard those materials.
>
> I think in terms of elementary, we do not have enough personnel do get everything done. Sometimes when we have a large number of small groups, we have to stagger our schedule so that we can have the staff to administer these exams.

Participant B

> When I became an assistant principal, I was moved into a school where there were a lot of test security issues. So my first experience with high-stakes test was answering questions regarding test security stemming from the previous test coordinator.

Participant C

> When we are talking about high-stakes testing in particular, you just have to make sure that you follow the proper process and procedures in reporting and recording things.

Participant D

> There is a lot of stopping and starting because there are a lot of other duties that happen throughout the school day. It is stressful, because although you go through the rules and do's and don'ts of testing . . . you really have to make sure that people understand what those rules are because you never won't people to . . .have to do a testing irregularity.

Participant E

> On the elementary level, my first year I was a test coordinator we had a state monitor, having a state monitor had me very worried because someone was actually scrutinizing every step of the way, from how the information entered the building to how it left the building to after testing.

Test coordinators' responses to high-stakes testing were similar to how they responded regarding their experiences as a test coordinator. Test coordinators' perceptions of high-stakes testing were not defined as it merely relates to views or opinions. Their perception of high-stakes tests was directly related to their own personal experiences. Thus, the collected responses gave an in depth view into the administration of a high-stakes test, and how it impacts the perception of test coordinators on high-stakes testing.

For instance, throughout the responses test coordinators' described high-stakes testing in terms of the actual process. Much detail was recounted regarding the process because the process lead to the test coordinator's perception of high-stakes testing. Inadvertently, factors that emerged were of security, resources, planning/organization, and fidelity. There seemed to be a connection between the emerging factors, as if one were dependent upon the other. For example, the perception that planning/organization were paramount to the success of a high-stakes testing program appeared in most

participant responses: *Participant B* ". . . I am at the front end of planning everything out. I have to get all the materials organized, and when that is done everything flows very smoothly." *Participant C* ". . . just having a plan. Make sure you are planning ahead." Planning and organization was the key to other factors that emerged such as security, resources, and fidelity. A successfully planned and organized testing environment minimized any issues with test security, increased the fidelity of test administration, and determined the necessity of resources.

Although a successful test administration required planning and organization, the amount of time allotted for this planning could be a week. Participant B stated, "prior to the tests I usually spend the entire weekend and several nights the week before," whereas, both Participants C and E stated that planning for high-stakes takes a "good week." The reason the planning and organization took a week was because the following had to be organized and planned: test materials, accommodation schedules, test security codes, test administrator schedules, etc. The participants had the following to say regarding the organization and planning:

Participant A

> . . . We have to meticulously go through each test and have the teacher sign-them out. Each test has to be signed out with the test numbers assigned with them. . . we are able to validate each test that was issued to them and sign-off on.

Participant B

> . . . Getting everything labeled all the pencils and all of the books in order. Sign-out records, the numbers have to match the books and so getting it in order takes a lot.

Participant C

> We strategically go through each one of the testing tubs checking for student accommodations. Organizing test for our students with accommodations takes time. I have to make sure that I have enough people to administer test to each one of those groups.

Participant D

> Unpacking the materials after you count. . . I try to keep the numbers in sequential order, then actually writing down a student name beside each number so that if you ever turn a booklet in and they say we didn't receive this booklet, I'll know.

Participant E

> Make sure that you had the space and proper accommodations and that you thought about everything from what would happen if a child got sick; to what would happened if there was an emergency. Making sure you have the proper and correct number of test booklets for students. . . having a staff meeting to explain their role in testing and what is expected of them.

The amount of time required for the administration of a high-stakes test takes slightly less time than the organization and planning of the test. Test coordinators estimated between 3 to 5 hours per subject area administered on testing day. This amount of time includes dissemination of testing materials to teachers, transitions, test administration, and collection/return of testing materials. Some of the test coordinators had the following to say regarding testing day:

Participant B

> The time of test administration actually depends upon the subject area of the tests. Reading usually lasts a little bit longer than science and social studies. But the tests for each day usually lasts about 2 hours for regular students and about time and half for small group students which is about 3 hours.

Participant C

> We start at 8:30. Teachers can get the tests as early as 7:30. The teacher should have picked up their tests by 8:15. They usually finish by 10:30 or 11:00 each day of testing.

Participant E

> Total with pre and post administration, I would say about 4 hours a day for each test and we had four or five test.

Test Coordinators' Perception of Computer-Based Testing

This section describes test coordinators' perceptions of computer-based testing prior to discussing the simulation results. Figure 5 shows the themes of security, emotions, and management also appeared in post-simulation responses. There was, however, a slight difference in the factors that were associated with the themes. Since the test no longer required the use of pencils and answer documents, erasure concerns were eliminated as a factor associated with security. A conducive testing environment was a new factor associated with test security. Test coordinators viewed the layout of the computer lab as an area of concern, especially with shortened test. Also, an increase in the number of purchased laptops would pose a security issue due to the lack of storage. The simulation evoked varying levels of emotions. Emotions ranged from excitement to fear and ambivalence. Factors associated with post-simulation processes remained the same. The description of these factors by test coordinators, however differed due to the computer administration.

Test coordinators' experiences with computer-based tests did not mirror those previously described. The level and variety of computer-based testing experiences

differed among the test coordinators. Test coordinators' experiences with computer-based testing ranged from diagnostic to high-stakes. Only one test coordinator had experience coordinating a high-stakes computer-based test; however, all test coordinators were aware of the online administration of the PARCC assessments. Test coordinators' experiences with computer-based testing were similar to those described in the previous section due to the emphasis on process. Still, the computer-based testing process required organizing, planning, and training, however, the magnitude to which this was done was less than the previous experience. The primary focus also shifted, from test security to resources (computers). The number of computers available to administer the test was limited, thus causing test coordinator's to create a variety of testing schedules (Appendix C).

Participant A

> However, I always thought that anytime students were given a diagnostic test that the time it takes to receive the exam results are so late in the summer that you do not have an opportunity to really carve out a plan for the student. So, I was really under the impression that for us going computer-adaptive . . . that the turnaround time for the results of these exams would be at a faster time. Also I thought a computer-adaptive test would protect the fidelity and security of the tests.

Participant B

> Actually, last spring I was part of a pilot program where we administered the high-stakes re-test on the computer. And that was actually a very good experience, mostly because I did not have all of the materials to organize and to distribute. I could get everything together in the database and uploaded the students and classes trained the teachers, I got a support system and trained them. And on the day of the program, everything just went very smooth.

Participant C

> I know that it (future assessments) is going to be where the students are testing online in each subject area. It will not just be multiple-choice, but students will have to get some written responses or they will have to explain their thinking processes.

Participant D

> My part was to make sure that the space was available and scheduling for the lab, to make sure that everyone has the opportunity to bring their classes in to get their test done in a timely manner. But we found out that we have to follow those specific directions. We found out that if we don't save, then it's not going through. It's like if we have kids in there for an hour for a test, but if we don't click save then it's like that didn't do anything.

Participant E

> In administering some of the benchmark and diagnostics to kids this past year, it was easier in terms of the logistics, the only problem or problems was what if the computer breaks down and there were issues some times where it was not doing what it was supposed to do. With technology sometimes, that will happen.

Test coordinators' perceptions of computer-based testing varied. Perceptions were formed based upon the test coordinators' experience and/or knowledge of computer-based testing. Since test coordinators experience and knowledge of computer-based testing varies, so did their perceptions. Factors that emerged were associated with time, security, resources, and training. Perceptions that stemmed from these factors were the rapidness of results, efficiency of the process, and using technology.

Several key words emerged from test coordinators' perceptions of high-stakes testing on the computer. These key words were: training, scheduling, and resources. Although, the words appeared in previous responses, they took on a different meaning

when discussed within the context of computer-based testing. For example, test coordinator's expressed the need for more specific training for high-stakes computer-based testing than paper and pencil testing. As in the case of high-stakes testing, resources were also in reference to personnel. For instance, the number of staff members required as part of compliance in test administration. However, resources in regards to computer-based testing refer to the number of computers needed to administer the test.

Participant A

> I would like to have more than one staff person to participate in a strong professional development or training to acclimate them to how to administer/implement a computer-based testing.

Participant B

> I knew exactly how many computers I had to work with, so I scheduled the students according to that. So, it was not just a morning testing period. We had a group in the morning and a group in the afternoon. It worked out very well. We had approximately 30 computers. It was not an issue in monitoring students . . . I had three test administrators in the lab at a time.

Participant C

> We need to make sure all the classes rotate through the computer lab. We have 28-30 computers with our largest class size around 26 students. The computer lab is closed on the days for testing. A schedule is created for those classes with each class rotating through the schedule each day.

Participant D

> I had to schedule the computer lab, to make sure that everyone has the opportunity to bring their classes in to the lab to get their test done in a timely manner. We have one computer lab, but we have 140 laptops. So, we were able to schedule between the laptops usage and being departmentalized helps out also. For the math, the math test, the math teacher got the laptops and used those in the classroom for one grade level. Then the other grade levels were able to run them through the computer lab. I had to do all three grade levels at one time because we do not have enough computers. Then the window, we have a week to do just one subject, but if we have to do all subjects within a week and three grade levels, I am really worried about that.

When asked about their perceptions of computer-based testing, many of the test coordinators had positive perceptions based upon their experiences. Most of test coordinators' experience with computer-based testing stemmed from administering the District's Benchmark Assessments, (one exception for high-stakes re-test) on the computer. In Appendix B, test coordinators provided evidence of how classes were scheduled in the computer lab for a benchmark assessment. Overall, the perception of test coordinator's to computer-based testing was positive. There were logistical issues regarding scheduling, for instance there were not any emotions/feelings of anxiety, stress, or feeling overwhelmed. When asked about their perceptions of administering a high-stakes test on the computer, the responses were slightly different. Test coordinators had the following perceptions of administering a high-stakes test on the computer:

Test Coordinators' Perceptions of Computer-Adaptive Testing

This section describes test coordinators' perceptions of computer-based testing after discussing the simulation results. Simulation results displayed in the earlier sections of Chapter Four were used to elicit a response from test coordinators concerning the administration of a high-stakes computer-based test. Elicited responses

varied among test coordinators. The process of administering a high-stakes CBT was still the primary focus for test coordinators. Test security and resources were not mentioned to the extent of the previous sections. Test coordinators' perceptions of CCT and VL-CAT varied from ambivalence to excitement. The ranges of perceptions were due to test coordinators' prior knowledge, understanding, and experience with CAT.

Even though increased test efficiency and precision were evident from the simulation results, it did not eliminate concerns test coordinators had in reference to high-stakes CBT's. More so, the efficiency and precision of CAT brought concerns related to the testing process not seen in previous discussions. One concern involved students who completed the test after only a few administered items. Other concerns included the level of computer knowledge of the student as well only using CCT and VL-CAT for diagnostic purposes.

Participant A

> I am a little ambivalent. I think in some terms in would be good, in saving money. The printing costs at time depending on whether it's diagnostic or high-stakes. Kids will have some pluses because they use the technology.

Participant B

> It is going to be a learning curve or is a learning curving for the teachers and the students are not accustomed to. I think it is a move in the right direction. I feel that as a society, everything is moving towards technology. It is preparing our students to think globally. So, we are just preparing our students for what lies ahead. I think that the process for giving tests will be a lot smoother for when we go to computer-based. I am excited about it.

Participant C

> I don't think it's a bad thing because I think it is a better perception of what the kids can actually do. We can't do that to kids and we can't do that to teachers. My only concern is that kids have enough practice in taking this type of test, that they can be successful in taking any type of computer test. Because if they are not taught or given the specific training of what they should be doing when taking this specific type of test then they are not going to perform.

Participant D

> So, it's still those directions that are very, very important that everyone follows. I don't feel good about. Let me explain that. The students are going to need practice on what is expected of them for high stakes testing. Those are some of the things I think about when I think about computer based testing, I don't know if I am thinking of it in a little too much detail of what they are asking kids to do

Participant E

> I am little apprehensive because if there is a glitch and under this whole atmosphere of test security and cheating. . . if there is a problem a teacher or test administrator may have a bit of fear that if I am assisting the student with doing anything to the tests. I can foresee some issues with that in terms of student and teacher readiness.

The response of test coordinators after being shown the simulation results were similar to the analysis of their perceptions of computer-based testing. Test coordinators' perceptions were formed based upon their own experiences and knowledge of the computer-based testing and computer-adaptive testing. Factors that emerged after reviewing the results from the computer-adaptive test were similar if not the same as far as resources and the use of technology.

Perceptions of computer-adaptive testing ranged from ambivalence to excitement. Time was a factor that emerged as it relates to students and the computerized

classification test. Some of the test coordinators viewed the shortened time as a possible classroom management problem for students who finished the test with only five questions. Test coordinators' perceived the shortened length for some students would cause curiosity with students who did not finish as quickly. Test coordinators' perception of test ending at varying lengths was mixed. Some test coordinators had a better perception of the test that terminated after the change in standard error reached .005 or less due to an even distribution of scores as compared to the computerized classification test.

Test coordinators' perception of computer-adaptive test without the variable-termination points varied due to their experience and knowledge of the test. Overall, the perception was that computer-adaptive test would cause less frustration in students because questions administered to students were specific for that student's ability level. Test coordinators also perceived computer-adaptive tests as a better measure of what students know.

Perceptions of using computer-adaptive test for high-stakes testing were questionable among most test coordinators. Test coordinators perceptions of high-stakes computer-adaptive tests were based upon issues of resources, scheduling and ease of computer use. Test coordinators stated that the amount of resources would be a challenge for schools due to the limited number of computers located within a building. The limited numbers of computer resources posed further problems with scheduling. To schedule all students within one computer-lab for several test administrations posed a challenge to test coordinators unless there was flexibility in test administration. By

flexibility, test coordinators noted that extending the testing window would allow them flexibility to schedule all students within the computer-lab.

Participant A

> I guess it's good and bad. It seems like it needs to be a little less cumbersome and confusing. I think it would be a good tool to have as far as diagnostics for kids. It would be a good tool to show what type of remediation the students would need to have. I do not feel as comfortable for using it for a high stakes tests as I would diagnostically.

Participant B

> Looking at this makes me even more excited. Also, it is telling me that for some of our students, it will decrease their frustration level because it will gauge their level of precision and kind of you know tweak the questions for their level. A lot of our students get really bored and lose their focus with the length of the tests; so they don't score as well as they should have.

Participant C

> If you are going to end after kid say takes 5 or 10 consecutive questions in a row. I think it would be a problem. You know it would be a distraction if a student finished. You are not going to leave them there because they would start acting out. We would have to remove them from the testing environment when they finished. If you are going to do it that way we would have to do it in a smaller setting.

Participant D

> I can see how it could be a benefit to the child instead of having them struggle through the same type of question on a paper pencil test, if the computer is going to adapt to the specific learner depending on the person taking the test, I can see how that's going to be a benefit to them. But I think that will be great because I think that when you get on a test and you see that a question is easy you are like it builds your confidence, so it will be a confidence booster for our children. They can feel that they are successful, instead of being all over the place with high and low questions with the level of rigor, they can feel a little bit more success.

Participant E

> I really think this concept is awesome. We just really need to build in a communication piece so that all stakeholders have a true understanding of the purpose of it. These are real results from my kids here . . . I think it's great. But it is how we use it for interpretation that will be a major piece; it will be a major paradigm shift as well.

Figure 7. Angle 1 Computer Lab

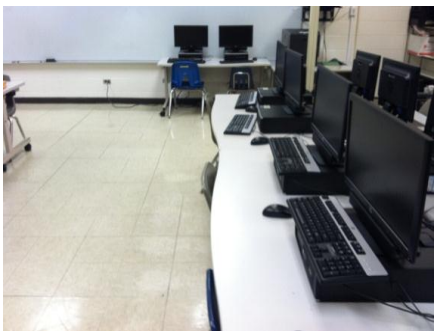


Figure 8. Angle 2 Computer Lab

Figure 9. Angle 3 Computer Lab

Figure 10. Angle 4 Computer Lab

Figures 7-10

Figures *7-10* are evidence of the test coordinators' description of the computer labs. Each computer lab contains thirty desktop computers. Several of the participants perceived resources as a challenge in administering high-stakes test on the computer. Compromising test security due to the proximity of computers to each other was one challenge. Another challenge was the limited number of computer resources. Although, VL-CAT and CCT would decrease the number of items administered to different

students, this seemed to pose further challenges. Test coordinators stated, that examinees would be distracted due to other examinees completing the assessment early. Test coordinators also mentioned that the schools had one computer-lab containing approximately 30 computers.

Schedules

Scheduling the computer lab for students to use for testing was discussed in several of the interviews. The following documents are schedules submitted by the interviewee for District Benchmark Assessments administered on the computer (Appendix A). The schedule describes who is scheduled for the computer lab, allotted time, day, and subject. Test coordinators provided schedules that were used to administer a district-wide computer-based test. The district assessment included the following subject areas: math, reading, language arts, science and social studies. Math assessments were administered to two grades online. All other subject area tests were administered via paper and pencil. As evident in the attached schedules, administering an online assessment for two grade levels and subject takes approximately a week. This is dependent upon the size of the grade levels as well as the number of available computers. The attached schedules also provide evidence of the amount of scheduling and organization that has to take place in order to administer a test.

CHAPTER 5

DISCUSSION

This chapter of the dissertation includes an overview of the study, a summary of the findings, draws major conclusions, and makes recommendations for future research. A detailed discussion of the findings is given in chapter 4.

The next generation of assessments is on the horizon. School districts across the United States will have to administer English Language Arts and Mathematics assessments online by the end of the 2014-2015 school year. This Next Generation of Assessments was part of the federal Race to the Top Assessment Program of 2009. As part of this federally funded initiative, each state joined one of two Consortia, Partnership for Assessment of Readiness for College and Careers (PARCC) or Smarter Balanced Assessment Consortium (SBAC) in order to build the framework for the assessment system (Center for K-12 Assessment & Performance Management, 2012).

PARCC and SBAC both require the administration of online assessments for students in k-12 public education. However, PARCC made some concessions regarding students in grades 3-5 and students with accommodations. These students will be allowed a paper and pencil administration "until studies confirm that students in these grades are ready for computer-based assessments (Center for K-12 Assessment & Performance Management, p. 16)". SBAC will offer paper and pencil administration for 3 years to offer school districts flexibility in the transition to computer-based assessments (Center for K-12 Assessment & Performance Management, 2012).

Although both Consortia require the use of an online test administration, only SBAC computer-version will be adaptive. The adaptive test tailors the difficulty level of each item from the student's response to the previously answered item. Adaptive test allows for a more efficient and secure way to measure student ability due to fewer test items administered to students, as well as the requirement of large item banks (Smarter Balanced Assessment Consortia, 2012).

Prior to the next generation of assessments and the use of online assessments, there was much debate regarding the use of computer-based assessments for high-stakes testing. Key stakeholders such as policymakers, test developers, and school district leaders all had a voice in the debate over computer-based testing. Policymakers' main point of contention was the idea that adaptive tests tested students off grade level. Test developers contended that the only major issue with computer-based assessments, specifically adaptive test, was the requirement of large item banks. State and District Level school officials questioned schools' readiness for computer-based testing as it relates to infrastructure and computer availability.

The purpose of this study was to investigate the degree of the efficiency and precision of computer adaptive and computer classification tests compared to those of paper and pencil benchmark tests and explore how the simulation results changed the perceptions of school test coordinators on high-stakes computer-based testing. The school test coordinators' take readers on a journey of testing in their perspective schools. The results from the CATSim Program were used to determine whether or not school test coordinators' views regarding computer-based testing would shift.

School test coordinators provided supporting evidence on their viewpoints of computer-based testing. The research questions guiding this case study were 1) Are the efficiency and precision of computer-adaptive tests or computer classification tests equivalent or superior to those of paper and pencil benchmark tests? (2) What are test coordinators' perceptions of administering high-stakes tests on the computer? (3) To what extent, if any, did computer simulation results elicit a change in test coordinators' perceptions of administering high-stakes tests on the computer?

Within the case study both quantitative and qualitative methods were used to collect the data required to answer the research questions. To address the efficiency and precision of CAT and CCT, item responses from the benchmark assessment were collected from the schools of participating test coordinators. The results from the first research question were then used to elicit responses from the school test coordinators. This approach to the study gave a comprehensive understanding of test coordinators' perceptions of high-stakes testing, computer-based testing, high-stakes computer-based testing, and subsequently computer-adaptive testing. Five in person interviews were conducted with test coordinators at each school. Each participant was asked guiding questions based upon a phenomenological framework. The rationale for this framework was to answer the research questions through the reflection of the participants' experiences as a school test coordinator. Perceptions of testing, high-stakes testing, and computer-based testing were formed as participants reflected on their experiences.

Since test coordinators had limited if any experience with computer adaptive test, results from a computer-adaptive simulation program were used during the

interview. Exposure to the simulation results stimulated school test coordinators to form a perception of how this type of adaptive test could impact their school's testing program. In addition to the semi-structured interviews, test coordinators were asked to provide supporting evidence to their responses.

Summary of Findings

The results from the simulations were used as part of the interview with the test coordinators. It was clear from the simulation results that CAT was a more efficient test compared to paper and pencil. The paper and pencil test required students to take an hour for administration whereas, VL-CAT and CCT test administration cut the testing time significantly. The amount of questions administered to students using the VL-CAT was reduced by seven items for half of the students. For CCT, two hundred and six students were administered only five items.

Test precision was measured in regards to the size of the standard error measure. How well did CAT precisely measure the theta level of the examinee? A large standard error denoted less test precision. In both simulations, VL-CAT and CCT, the mean standard error was greater than the full bank thetas. This is expected, because the longer tests are expected to have more precision. The relationship between full bank thetas and CAT thetas were strong and positively correlated in both VL-CAT and CCT. However, the correlation of the standard errors of full bank and CAT thetas differed within the standard errors of each test. CCT standard errors theta had a stronger positive correlation than VL-CAT thetas. The standard error correlation of full length test thetas and VL-CAT theta's although

positive were much weaker than CCT. This is explained by the variance of the VL-CAT, it took longer for VL-CAT to determine the precision level.

The RMSD compared the theta estimates for VL-CAT and CCT, the shorter tests, to those of the full length test. Therefore, the smaller RMSD is desirable. It was unique to find that for this case study, the RMSD for CCT was smaller than that of CCT, especially since CCT was the shorter test. Specifically, CCT achieved better efficiency and precision than VL-CAT for this study. Overall, the simulation program demonstrated to research participants the efficiency and precision of CAT. Although, the efficiency measured by the number of items administered to examinees was clear to the participants, the extent of their understanding was unknown. In order to address the last research questions, test coordinators reflected upon their experiences with high-stakes testing, and computer-based testing. As a result, the following themes emerged: security, emotions, and management.

The following information was revealed regarding coordinating a testing program: (a) it can evoke certain emotions/feelings such as anxiety, stress, or frustration, (b) it requires knowledge and patience, and (c) it requires security, organization/planning, time, and fidelity. Test coordinators' experiences with coordinating high-stakes tests were similar to their responses in coordinating all tests. Test coordinators wanted to ensure that the process of administration was followed so the test would be administered with fidelity. To ensure the fidelity of the test, test security had to be ensured which would only occur by planning and organization. In summary, a test coordinators experience with testing in general including high-stakes testing reveals that it is a process that requires time, patience, and organizational skills if it is to be implemented with fidelity.

Data collected on the research question, *What are test coordinators' perceptions of a computer-based testing program?* revealed that test coordinators' perceptions of computer-based testing were formed from their experience and/or prior knowledge of computer-based testing. Test coordinators formed a variety of perceptions on computer-based testing such as (a) test results are reported faster (b) a higher level of test security/fidelity (c) less time is required to organize test materials (d) shorter test (e) uses artificial intelligence to score examinee responses (f) technical issues (saving, computer breaks down) that could result in score loss, and (g) easier logistics.

To address the research question, *To what extent, if any, did computer simulation results elicit a change in test coordinators' perceptions of a computer-based testing program?* test coordinators were given an explanation of computer-adaptive testing and shown the simulated results for a CCT or CAT. Test coordinators had the following perceptions of adaptive testing once the simulation results were discussed: (a) increase knowledge and understanding of computer-adaptive test to all stakeholders (b) use as a diagnostic tool for students (c) it will lower the frustration level of students (d) it will help gauge where students are with their learning (e) increase the confidence level of students, and (f) increase student focus on the test.

The purpose of research question *to what extent, if any, did computer simulation results elicit a change in test coordinators' perceptions of a computer-based testing program?* was to determine how simulation results impact test coordinators' responses as it relates to the emerging themes. Prior to showing test coordinators the results of the simulation, most of their perceptions of computer-based testing were due to their current experiences with testing in general. Test coordinators

experiences with high-stakes test using paper and pencil were stated as being an "overwhelming" and sometimes stressful process if test coordinators were not organized. Most test coordinators welcomed the idea of using computers for testing, however, they were quick to note some of the challenges that come with this type of testing. The most prevalent challenges were those regarding resources.

Test coordinators expressed that computer labs only have 30 computers with schools with more than 30 students. The logistics of administering a test online presented a unique challenge. Familiarity with the computer by the student as well as the teacher was also a topic that appeared. Some students are more comfortable using technology than other students, whereas all students would need to learn test taking strategies for computer administrations. Although test security was an area of focus for paper and pencil test, it did not present a problem for computer-based tests.

After test coordinators were shown the simulated results from CCT and CAT, did their perceptions of computer-based testing change? Overall, the response to CCT and CAT were positive. Test coordinators were amazed by the number of questions administered to students before a test would terminate. The perception of administering students a test that would "differentiate" the test item based upon students responses was received positively. However, test coordinators perceived the shorter tests as potentially problematic. The problem was associated with students having different stopping points as in the case of VL-CAT and CCT. Although varying the number of questions appeared to be a solution to scheduling students in the computer lab with 30 computers, test coordinators were concerned

about the perception of early termination. Test coordinators perceived students would be distracted if other students were completing the test prior to them finishing. Students who completed the test early would not be able to leave the testing environment.

In conclusion, the simulation did not change the perception of computer-based-testing. Shortened test were perceived as positive by test coordinators. However, the simulation did not change the perception of a high-stakes computer-based testing due to the ever-present challenges of scheduling and resources. Even though, shortened test would make more computers available at a faster rate, thus alleviating the problem of too few computers. Participant A, stated it best when asked about the use of CAT for high-stakes testing, "I think it would be a good tool to have as far as diagnostics . . . it would be a good tool to show what types of remediation the students would need to have."

Conclusion

To build sustainability of the next generation of assessments, it is imperative to involve all stakeholders in the process. A whole is only as good as its parts. The same is true in education. There are many parts to educating the whole child. Thus, in ensuring that the next generation of assessments is implemented with fidelity, everyone must be given a voice in the process. Policymakers, test developers, District leaders all have had input in the process, however the individuals who have the most impact are the voices we do not hear.

This case study revealed how the efficiency and precision of CCT and CAT were instrumental in forming the perceptions of test coordinators regarding high-stakes testing

on the computer. This study allowed insight into how a computer-based testing program would impact the school. The study also demonstrated the importance of a simulation research prior to making a computer-adaptive test or any other computer test operational. Furthermore, this study proved the importance of data triangulation. Not by just triangulating data through multiple qualitative data points, but more so, using quantitative data as well. This study used data from multiple sources such as photographs and documents to validate the responses of the interview participants.

In addition, adaptive test simulations were run using responses from students from the interviewee's school. Thus, making the results meaningful to the interviewees as well as providing data on test efficiency and precision in adaptive testing. If a simulation study was conducted without interviewing test coordinators, the results would have been interpreted from only one perspective---quantitatively. The use of qualitative research methods revealed profound evidence, that although CATs are efficient and precise, this is only one component to testing. This study proved that perceptions are based upon the lens in which it is viewed. Quantitatively speaking, if VL-CAT and CCT provided answers to test efficiency and precision, more examinees could be tested due to the higher level of efficiency and precision of the test. Qualitatively, test coordinators wondered what would happen to examinees as individuals completed tests earlier than others. If there is no standard time for everyone to finish, then how can a conducive testing environment be maintained for all students?

In summary, the results from this study encapsulated the importance of the case study design. Although, the findings of the study cannot be generalized to other school districts, the process can be replicated. The phenomenological framework grasped test

coordinators' experiences which were unique to their school district. These experiences revealed that the perceptions of test coordinators on computer-based testing were different from psychometricians. For instance, issues of test security. Psychometricians view security for computer test in terms of item exposure, whereas test coordinators' viewed the security of a test as it relates to cheating.

There were several limitations to the study that were also unique to this study. For example, subject matter of the test, examinees with accommodations, and the 1.0 ability level used to categorize examinees as pass or fail on the CCT. The Social Studies test required examinees to read and understand text, which would impact the responses of examinees with limited reading ability. Special populations of examinees, ESOL or Special Needs, received testing accommodations that extended the testing time and allowed for test items to be read. The cutoff value for the CCT was selected by the researcher based on the ability estimates of the examinees. To conclude, the details of each case have an impact on the results of the study.

Recommendations

The implications for further research are limitless in the area of high-stakes computer-based testing in k-12 schools. Recommendations for additional research from the current study involve both qualitative and quantitative methods: (a) conduct additional post-hoc simulation studies using a variety of termination points (b) compare student scores from paper and pencil administration and computer-adaptive administration (c) examine types of questions most frequently administered to students (d) explore parent, teacher, student perceptions of computer-based tests, and (e) conduct a cost-benefit analysis of paper and pencil to computer-based tests. A final

recommendation is to continue to conduct research using simulations and the perspectives of test coordinators prior to the implementation of a computer-based testing program.

References

Achieve (2010, August). *On the road to implementation: Achieving the promise of the common core state standards.* Retrieved April 4, 2012 from http://www.achieve.org./files/FINAL-CCSSImplementationGuide.pdf

Babcock, B. & Weiss, D. (2009). *Termination criteria in computerized adaptive tests: Variable-length CATs are not biased.* In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on *Computerized Adaptive Testing.* Paper presented at the Realities of CAT Paper Session.

Broadfoot, P.M. (1979). *Assessment, schools and society*. Methuen: London, England.

Center for K-12 Assessment & Performance Management (2012). Coming Together to Raise Achievement. New Assessments for the Common Core State Standards. Educational Testing Center.

Cohen, A. & Wollack, J. (2006). *Test Administration, Security, Scoring, and Reporting.* Educational Measure, 4th Edition. Praeger: Westport, CT.

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process.* Thousand Oaks, CA: Sage.

Creswell, J.W. (2003). *Research design: Quantitative, qualitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

CTB-McGraw-Hill, LLC (2013). *Transitioning to online assessments—using technology to enhance learning.* Retrieved April 8, 2013 from http://ctbonlineassessments.com/eguide

Davey, T. (2011). *Practical considerations in computer-based testing.* Princeton, NJ: Educational Testing.

Davey, T. & Pitoniak, M. J. (2006). Designing computerized adaptive tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 543-573). Mahwah, NJ: Lawrence Erlbaum Associates.

de Beer, M. Visser, D. (1998). Comparability of the paper- and-pencil and computerized adaptive versions of the general scholastic aptitude test (GSAT) senior. *Journal of Psychology,* 28(1), 21-28.

Gipps, C. (1999). Socio-cultural perspective on assessment. In A. Iran-Nejad, & P.D. Pearson (Eds.), *Review of Research in Education.* 24, 355-392. Washington: American Educational Research Association.

Governor's Office of Student Achievement. (2011). *K-12 Public School Report Cards, 2011* [data file]. Available from The Governor's Office of Student Achievement Website, http://gosa.georgia.gov/contents-report-card

Grunwald, Assoc. (2010). *An open source platform for internet-based assessment: A report on education leaders' perceptions of online testing in an open source environment.* Retrieved November 8, 2010, from http://www.grunwald.com

Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park: Sage.

Harmon, D. J., (2010, June). *Multiple perspectives on computer adaptive testing for K-12 assessments.* Policy Implications from the federal perspective. U.S. Department of Education.

Horn, R. V., (2003). Computer adaptive tests and computer-based tests. *Phi Delta Kappan*. 567 & 630-631.

Impara, J., & Foster, D. (2006). Strategies to minimize test fraud. In S.M. Downing &

T.M. Haladyna (Eds.), *Handbook of test development* (pp.91-114). Mahwah, NJ: Lawrence Erlbaum.

Jones, P., Smith, R., & Talley, D.M. (2006). Developing test forms for small-scale achievement testing systems. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp.487-525). Mahwah, NJ: Lawrence Erlbaum.

Keng, L., McClarty, K.L., & Davis, L.L. (2008). Item-level comparative analysis of online and paper administration of the Texas assessment of knowledge and skills. *Applied Measurement in Education, 21(3),* 207-226.

Kingsbury, G.G., & Hauser, C. (2004, April). *Computer adaptive testing and no child left behind.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kolen, M. & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Koretz, D. & Hamilton, L. (2006). *Testing for Accountability in K-12.* Educational Measurement. Prager: Westport, CT.

Lazer, S., Mazzeo, J. Twing, J.S., Way, W.D., Camara, W., & Sweeney, K. (2010, May). *Thoughts on an assessment of common core assessments* (ETS, Pearson, & the College Board white paper). Princeton, NJ: Educational Testing Services.

Lazer, S., Mazzeo, J. Twing, J.S., Way, W.D., Camara, W., & Sweeney, K. (2010, Feb.). *Some considerations related to the use of adaptive testing for the common core assesments.* (ETS, Pearson, & the College Board white paper). Princeton, NJ: Educational Testing Services.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.*

Hillsdale, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (2006). Designing tests for pass-fail decisions: Using item response theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 575-595). Mahwah, NJ: Lawrence Erlbaum Associates.

Mazzeo, J., & Harvey, A.L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of literature* (College Board Rep. No. 88-8, ETS RR No. 88-21). Princeton, NJ: Educational Testing Service.

McCallin, R.C. (2006). Test administration. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp.625-652). Mahwah, NJ: Lawrence Erlbaum.

Mislevy, Robert (2006). *Cognitive Psychology and Educational Assessment.* Educational Measurement, 4th Edition. Prager: Westport, CT.

Moustakas, C. (1994). *Phenomenological research methods.* Thousand Oaks, CA: Sage.

No Child Left Behind, testing and flexibility. (2007). *The Washington Post*, November 27. Retrieved November 20, 2011 from http://www.publiceducation.org/nclb_articles/archive/20071127_NCLB.asp

Olson, L. (2003). Legal twist, digital turns: Computerized testing feels the impact of *No Child Left Behind*. *Education Week*, 22 (35), 11-15, 16.

Park, (2003). A test takers perspective. *Education Week*, 22(35), 15-16.

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer.

Poggio, J., Glasnapp, D.R., Yang, X., & Poggio, A.J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in

large scale state assessment program. *Journal of Technology, Learning, and Assessment,* 3(6).

Schmeiser, C.B. & Welch, C. J. (2006). *Test Development*. Educational Measurement. 4th Edition. Praeger Publishers. Westport, CT.

Smarter Balanced Assessment Consortium. Retrieved December 31, 2012 from www.smarterbalanced.org/smarter-balanced-assessments/computer-adaptive-testing/

Stake, R. E. (2005). *Qualitative case studies.* The Sage Handbook of Qualitative Research. 3rd Edition. Sage Publications. Thousand Oaks, CA.

State Educational Technology Directors Association (2011, June 22). *Technology requirements for large-scale computer-based and online assessment: Current status and issues.* Retrieved from http://www.setda.org./c/document_library/get_file?folderld=344&name=DLFE-1336.pdf

Texas Education Agency. (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests.* Retrieved April 4, 2012 from http://ritter.tea.state.tx.us/student.assessment/resources/techdigest/Technical_Reports/2008_literature_review_of_comparability_report.pdf.

Thompson, N. & Weiss, D. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation,* 16(1).

Trotter, A. (2003). A question of direction: *Adaptive* testing puts testing officials and experts at odds. *Education Week,* 22 (35), 17-21.

U.S. Department of Education, Office of Educational Technology. (2010). *Transforming American education: learning powered by technology* (ED-040CO-0040). Washington, DC: Education Publication Center.

Washington, Wayne (2013, July 22). Georgia decides against offering common core standardized test. *Atlanta Journal Constitution*, Retrieved July 22, 2013 from http://www.ajc.com/news/news/breaking-news/georgia-decides-against-offering-common-core-stand/nYzDr/

Way, W.D., Um, K., McClarty, K.L. (2007, April). *An evaluation of a matched samples method samples method for assessing the comparability of online and paper test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois.

Weiss, D., & Guyer, R. (2010). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul MN: Assessment Systems Corporation.

Wise, S.L., & Plake, B.S.(1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, 8(3), 5-10.

Yeh, S. (2006). Reforming federal testing policy to support teaching and learning. *Educational Policy*, 20(3), 495-524.

Schaffhauser, D. (2011). High-stakes online testing, T H E Journal, 38(6). Retrieved [data] from: http://thejournal.com/articles/2011/06/07/high-stakes-online-coming-soon.aspx

Yin, R.K. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks: Sage.

APPENDIXES

APPENDIX A

INTERVIEW QUESTIONS

1. Take a few minutes and reflect on your experiences as the school test coordinator.
2. What feelings or emotions come to mind?
3. Can you tell me about your last experience coordinating a high-stakes test?
4. Is there anything about this experience that stands out?
5. How does the experience of administering high-stakes impact you?
6. Describe any experiences you had administering tests on computers?
7. Are you aware of the new common core assessments?
8. If so, what is your understanding of this new form of assessment?
9. What is your perspective of implementing high-stakes tests on the computer? Do you have any evidence to support your claims?
10. Are you aware that some Common Core Assessments will be computer-adaptive?
11. What is your understanding of a computer-adaptive test? Explain computer-adaptive testing to participants. Show results from simulation.
12. What are your perceptions of computer-adaptive testing?
13. Does your current understanding of computer-adaptive testing, change your perception of implementing a high-stakes test on the computer?
14. Why has your perception of computer-based testing changed or remained the same?

APPENDIX B

TEST SCHEDULES

District Benchmark Test Administration Schedule	**Monday November 12th**	**Tuesday November 13th**	**Wednesday November 14th**	**Thursday November 15th**	**Friday November 16th**	**Monday November 26th**
Subject Tested 8:10am-9:20am	Math-**4th grade only**	Reading	Language Arts	Science	Social Studies	Social Studies Makeups
Makeups 10:00am-11:10am	Math- 4th grade Tardy Students	Math -4th grade Absent Students Reading-Tardy Students	Reading – Absent Students ELA- Tardy Students	ELA- Absent Students Science-Tardy Students	Science-Absent Students Social Studies-Tardy Students	
Subject/Class Tested 11:00am-12:15pm	Math-	Math- ***Makeups-***	Math- ***Makeups-***	Math- ***Makeups-***	***Makeups-***	
Subject/Class Tested 12:45pm-2:00pm	Math-	Math- ***Makeups-***	Math- **Makeups-**	Math- ***Makeups-***	***Makeups-***	

Computer Based Testing Specifics

Third and Fifth grade students will take the Math Assessment in the Computer lab during the scheduled time.

3rd Grade - 8:15-9:45	**Monday, Nov. 12th**	**Tuesday, Nov. 13th**	**Wednesday, Nov. 14th**	**Thursday, Nov. 15th**	**Friday, Nov. 16th**
Computer Lab – Math Online	(Math)	(Math)	(Math)	(Math)	Make Ups….
Paper/pencil	(R/ELA)	(Science)	Soc.St.)	(R/ELA)	Make Ups….
Paper/pencil	(R/ELA)	(Science)	(Soc.St.)	(Science)	Make Ups…
Paper/pencil	(R/ELA)	(Science)	(Soc.St.)	(Soc.St.)	Make Ups…
4th Grade - 9:55-11:25 Computer Lab- Math Online	(Math)	(Math)	(Math)	(R/ELA) (paper & pencil)	Make Ups…
Paper/pencil	(R/ELA)	(Sci.)	(S.S.)	(S.S.)	Make Ups…
Paper/pencil	(R/ELA)	(Sci.)	(S.S.)	(Sci.)	Make Ups….
	5th Grade…In the classroom **paper and pencil** for Reading and ELA only.	**5th Grade Only…<u>Online</u> using laptops and desktops in the classroom**	**5th Grade Only…<u>Online</u> using laptops and desktops in the classroom**	**5th Grade Only…<u>Online</u> using laptops and desktops in the classroom**	Make ups…
5th Grade 9:30 – 11:00	(R/ELA)	(Math online)	(Science online)	(Social Studies online)	Make ups…
	(R/ELA)	(Math online)	(Science online)	(Social Studies online)	Make ups…
	(R/ELA)	(Math online)	(Science online)	(Social Studies online)	Make ups…

Elementary Testing Updates

November 2012

Benchmark Testing (3rd ,4th, and 5th) Nov. 13, 14, 15, and 27th Lab Closed Nov. 14th and 15th

Mon	Tue	Wed	Thu	Fri	Sat	Sun
			1	**2**	**3**	**4**
5	**6**	**7**	**8**	**9**	**10**	**11**
12	**13** Benchmark Test Science 3rd ,4th ,5th 8:30-10:30	**14** Benchmark Test Math 3rd Online Math 4th Paper Social Studies 5th 8:30-10:30	**15** Benchmark Test Math 5th Online Social Studies 3rd Social Studies 4th 8:30-10:30	**16** Make-up Testing	**17**	**18**
19	**20**	**21**	**22** Thanksgiving Day	**23**	**24**	**25**
26	**27** Benchmark Test Reading/ELA 3th-5th 8:30-10:30	**28**	**29** Make-up Testing	**30**		

APPENDIX C

VL-CAT AND CCT THETA INFORMATION AND SEM

Theta	Information	SEM
-3.00	0.167	2.4455
-2.95	0.182	2.3445
-2.90	0.198	2.2478
-2.85	0.215	2.1551
-2.80	0.234	2.0663
-2.75	0.255	1.9812
-2.70	0.277	1.8997
-2.65	0.301	1.8217
-2.60	0.328	1.7469
-2.55	0.356	1.6753
-2.50	0.387	1.6067
-2.45	0.421	1.5410
-2.40	0.458	1.4781
-2.35	0.497	1.4179
-2.30	0.541	1.3602
-2.25	0.587	1.3050
-2.20	0.638	1.2521
-2.15	0.693	1.2015
-2.10	0.752	1.1531
-2.05	0.816	1.1067
-2.00	0.886	1.0624
-1.95	0.961	1.0199
-1.90	1.043	0.9793
-1.85	1.131	0.9405
-1.80	1.225	0.9034
-1.75	1.328	0.8678
-1.70	1.438	0.8339
-1.65	1.557	0.8014
-1.60	1.685	0.7704
-1.55	1.822	0.7408
-1.50	1.970	0.7125
-1.45	2.128	0.6855
-1.40	2.298	0.6597
-1.35	2.479	0.6351
-1.30	2.673	0.6116
-1.25	2.880	0.5893
-1.20	3.100	0.5679
-1.15	3.335	0.5476
-1.10	3.584	0.5282

-1.05	3.847	0.5098
-1.00	4.126	0.4923
-0.95	4.421	0.4756
-0.90	4.730	0.4598
-0.85	5.056	0.4447
-0.80	5.397	0.4305
-0.75	5.753	0.4169
-0.70	6.124	0.4041
-0.65	6.510	0.3919
-0.60	6.909	0.3805
-0.55	7.321	0.3696
-0.50	7.745	0.3593
-0.45	8.179	0.3497
-0.40	8.622	0.3406
-0.35	9.073	0.3320
-0.30	9.530	0.3239
-0.25	9.990	0.3164
-0.20	10.451	0.3093
-0.15	10.912	0.3027
-0.10	11.369	0.2966
-0.05	11.820	0.2909
-0.00	12.262	0.2856
0.05	12.693	0.2807
0.10	13.109	0.2762
0.15	13.507	0.2721
0.20	13.884	0.2684
0.25	14.237	0.2650
0.30	14.564	0.2620
0.35	14.860	0.2594
0.40	15.125	0.2571
0.45	15.354	0.2552
0.50	15.545	0.2536
0.55	15.697	0.2524
0.60	15.807	0.2515
0.65	15.874	0.2510
0.70	15.896	0.2508
0.75	15.874	0.2510
0.80	15.807	0.2515
0.85	15.695	0.2524
0.90	15.539	0.2537
0.95	15.341	0.2553
1.00	15.101	0.2573
1.05	14.823	0.2597
1.10	14.508	0.2625
1.15	14.159	0.2658
1.20	13.781	0.2694

1.25	13.375	0.2734
1.30	12.947	0.2779
1.35	12.498	0.2829
1.40	12.034	0.2883
1.45	11.558	0.2941
1.50	11.074	0.3005
1.55	10.584	0.3074
1.60	10.093	0.3148
1.65	9.603	0.3227
1.70	9.118	0.3312
1.75	8.639	0.3402
1.80	8.170	0.3499
1.85	7.712	0.3601
1.90	7.266	0.3710
1.95	6.835	0.3825
2.00	6.419	0.3947
2.05	6.019	0.4076
2.10	5.636	0.4212
2.15	5.270	0.4356
2.20	4.922	0.4507
2.25	4.591	0.4667
2.30	4.278	0.4835
2.35	3.982	0.5011
2.40	3.703	0.5197
2.45	3.440	0.5391
2.50	3.193	0.5596
2.55	2.962	0.5810
2.60	2.745	0.6035
2.65	2.543	0.6271
2.70	2.354	0.6518
2.75	2.177	0.6777
2.80	2.013	0.7048
2.85	1.860	0.7332
2.90	1.718	0.7629
2.95	1.586	0.7940
3.00	1.464	0.8265

APPENDIX D

Software Programs

Software programs used in the study are available for purchase at the following addresses:

XCalibre 4

Assessment Systems Corporation
6053 Hudson Road, Suite 345
Woodbury, MN. 55125
http://www.assess.com/

CATSIM

Assessment Systems Corporation
6053 Hudson Road, Suite 345
Woodbury, MN. 55125
http://www.assess.com/