FACILITATING VARIABLE-LENGTH COMPUTERIZED CLASSIFICATION TESTING VIA AUTOMATIC RACING CALIBRATION HEURISTICS

Andrew F. Barrett

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements for

the degree

Doctor of Philosophy

in the School of Education,

Indiana University

April 2015

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Doctoral Committee

_____

Theodore W. Frick, PhD.

_____

Thomas A. Brush, PhD.

_____

Charles M. Reigeluth, PhD.

_____

Leslie A. Rutkowski, PhD.

December 1, 2014

I dedicate this dissertation to my family. I would like to thank my parents, Fred and Lois Barrett, for giving me a foundation of unwavering love and encouragement on which to build, and my brother, Adam Barrett, for being there for me during both fun and challenging times. This dissertation is also dedicated to my wife, Klaudia Papp, for being an amazing woman, a true partner in life and a source of support, love, and inspiration, and to our son, Adrian, for the joy and meaning he brings to our lives.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Theodore Frick, Dr. Thomas Brush, Dr. Charles Reigeluth, and Dr. Leslie Rutkowski for serving as members of my research committee. Thank you for sharing your valuable expertise, insight, and feedback. Your contributions have improved this dissertation and will positively impact my future scholarship.

I would also like to thank numerous Indiana University staff and fellow students for helping me throughout my journey to get the dissertation done. The Dissertation Support group led by Dr. Paul Toth and the MAPSAT research group led by Dr. Frick both provided timely and thoughtful scholarly assistance, guidance, and encouragement.

I was very fortunate to have had the support of good friends from the beginning of my doctoral studies. Thank you Jamison Judd, Shane DeMars, and Rose Benedicks for being there to enjoy the good times and for the help you gave when life got bumpy.

In the end, two extraordinary families made all the difference. I will forever be grateful to Craig Howard, Shoko Furuya Howard, and Erena Furuya Howard, as well as Rod Myers, Juliana Tagliaferri, and Miles. Thank you Craig and Rod for the direct contributions you made to this dissertation, for supporting my growth as a scholar, and, most importantly, for our friendship. Thank you Shoko and Juliana for sharing your kindness, wisdom, and generosity and for consistently steering Craig, Rod, and I in the right direction.

Two people deserve special acknowledgement. First, I would like to thank Dr. Frick for his steadfast support, excellent mentorship, and for encouraging the rich and collaborative scholarly environment that have benefited so many. Finally, I thank my wonderful wife, Klaudia Papp, for her love, understanding, and dedication. I could not have done it without you.

Andrew Frederick Barrett

FACILITATING VARIABLE-LENGTH COMPUTERIZED CLASSIFICATION TESTING VIA AUTOMATIC RACING CALIBRATION HEURISTICS

Computer Adaptive Tests (CATs) have been used successfully with standardized tests. However, CATs are rarely practical for assessment in instructional contexts, because large numbers of examinees are required *a priori* to calibrate items using item response theory (IRT). Computerized Classification Tests (CCTs) provide a practical alternative to IRT-based CATs. CCTs show promise for instructional contexts, since many fewer examinees are required for item parameter estimation. However, there is a paucity of clear guidelines indicating when items are sufficiently calibrated in CCTs.

Is there an efficient and accurate CCT algorithm which can estimate item parameters adaptively? Automatic Racing Calibration Heuristics (ARCH) was invented as a new CCT method and was empirically evaluated in two studies.

Monte Carlo simulations were run on previous administrations of a computer literacy test, consisting of 85 items answered by 104 examinees. Simulations resulted in determination of thresholds needed by the ARCH method for parameter estimates. These thresholds were subsequently used in 50 sets of computer simulations in order to compare accuracy and efficiency of ARCH with the sequential probability ratio test (SPRT) and with an enhanced method called EXSPRT. In the second study, 5,729 examinees took an online plagiarism test, where ARCH was implemented in parallel with SPRT and EXSPRT for comparison.

Results indicated that new statistics were needed by ARCH to establish thresholds and to determine when ARCH could begin. The ARCH method resulted in test lengths significantly

shorter than SPRT, and slightly longer than EXSPRT without sacrificing accuracy of classification of examinees as masters and nonmasters.

This research was the first of its kind in evaluating the ARCH method. ARCH appears to be a viable CCT method, which could be particularly useful in massively open online courses (MOOCs). Additional studies with different test content and contexts are needed.

_____

Theodore W. Frick, PhD.

_____

Thomas A. Brush, PhD.

_____

Charles M. Reigeluth, PhD.

_____

Leslie A. Rutkowski, PhD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I. INTRODUCTION

1.1 Background

Pressure is mounting on educators to better prepare learners to succeed in an increasingly knowledge driven economy (Gardner et al., 1983; Mourshed et al., 2010). A fundamental component common to many proposed strategies for improving the effectiveness of education (e.g. Reigeluth et al., 2008; Christensen et al. 2008; Collins & Halverson, 2009) involves ensuring information about the current state of an individual learner's knowledge informs and shapes educational decision-making. A lack of information about learner knowledge can hinder associated educational improvement efforts (Popham, 2003). Consequently, the extent to which an educator may access timely and accurate information about an individual learner's knowledge is a critical component of many efforts to improve the effectiveness of education.

Educator access to information about learner knowledge depends on the availability of two, often scarce, resources: time and quality assessments. Attempting to locate an appropriate assessment, its administration, and associated grading takes precious time. If a suitable assessment is not available then a choice must be made between creating one and abandoning the assessment effort. Given that educators often have neither the time nor the knowledge required to create quality assessments (Crooks, 1988; Black, 1993; Black & Wiliam, 1998), the choice between no and poor assessment represents a no-win situation. Educators need easy access to existing quality assessments that minimize time demands placed on both learners and educators in order to facilitate educator access to information about learner knowledge and to support associated efforts to improve education.

Trends in three areas may contribute to making high quality and efficient assessments increasingly available to educators in the not too distant future: Open Educational Resources (OERs), Computer Adaptive Testing (CAT), and mobile computing. OERs are defined as "digitised materials offered freely and openly for educators, students and self-learners to use and reuse for teaching, learning and research" (Hylén & Schuller, 2007). The OER movement provides the paradigm for increased creation, distribution, and use of educational resources such as digital assessments. However, access is only part of the problem – assessments must also be high quality and efficient.

CAT is a set of approaches whose primary aim is to dramatically improve test efficiency without compromising test validity and reliability (Thompson, 2007). Use of CAT by educators could reduce the class time necessary to administer an assessment. CAT approaches typically include procedures for examining the quality and validity of test items to address test quality. CAT depends on automating scoring of examinee responses and, therefore, is only appropriate for assessing specific types of learning. However, when CAT is appropriate, it could further reduce time demands placed on educators by automating grading.

The computing resources required for CAT, both from the perspective of the test administrator and that of examinees, is increasing available due to the growing access to computing resources, particularly due to advances in mobile computing (Triantafillou, Georgiadou, & Economides, 2008). Mobile computing involves new types of devices and a different approaches to software. Devices such as net-books, tablet computers, and smartphones can provide Internet access and substantial computing resources at a fraction of the cost of traditional desktops or laptops making them particularly attractive for use in

educational contexts. Free or low cost applications available to a variety of devices via a web browser (e.g. Google Docs) or app stores are increasingly competing with often-expensive software (e.g. Microsoft Office). While the application logic of CAT can be far from simple, it is not too complex to be delivered via this new approach to software on mobile devices.

1.2 Problem

Unfortunately, educators currently do not have open access to a wide range of assessments that apply CAT approaches to efficiently and accurately reveal what learners know. Furthermore, it is unlikely that they will have this access in the near future unless a critical problem is dealt with. If this problem is not addressed then an educator's lack of information about what their learners know will threaten the success of many efforts to improve education and better prepare learners to succeed in an increasingly knowledge driven economy. The following will first detail the general problem before explaining the specific problem that was the focus of this research.

The general problem is that heavy resources requirements associated with creating assessments that use CAT approaches, particularly item calibration, makes CAT impractical in all but a few large-scale, high-stakes, and/or highly profitable contexts. CAT depends on the availability of software and hardware for test administration, a bank of test items, and specific information about items established during item calibration. While CAT software can be reused across many testing contexts, test items are far less versatile. Often test items are only applicable to a narrow set of learning objectives.

To make matters worse, arduous item calibration approaches yield results whose value may be questionable and always degrade with use. While only one individual may be able to create an item, item calibration can involve gathering responses from hundreds or

thousands of examinees *before* the item can be used in CAT. Further, large motivational differences have been shown to exist between examinees who participate in item calibration, an often low or no stakes context, and the examinees for which tests are being designed for, which raise questions about the validity of item calibration data collected (Wise & DeMars, 2005; Makransky, 2008). Finally, items and associated calibration data degrade with use. Item exposure to examinees via testing serves to increase the odds of the item becoming compromised and, therefore, less effective for assessing examinee ability.

Item calibration need not be quite so involved. CAT approaches that make classification decisions about examinee knowledge, called Variable-Length Computerized Classification Testing (VL-CCT), have been shown to be highly efficient and accurate while requiring a substantially less arduous calibration phase. Application of Wald's () Sequential Ratio Probability Test (SPRT) requires little, if any, item calibration and has been shown to make accurate classification decisions while cutting average test lengths to a fourth of the traditional full length tests (Frick, 1989). Frick (1992) demonstrated that a calibration phase involving as few as 25 examinees from each classification group responding to all items in an item-bank and a modified version of SPRT that incorporated expert systems reasoning (EXSPRT) enabled even more efficient classification testing without compromising accuracy.

However, it cannot be assumed that the 25 examinees per classification group guideline is always appropriate for sufficiently calibrating an item-bank to enable VL-CCT testing. The guideline is based on only two sets of test data and does not factor in key details about the items being calibrated (e.g. item discrimination and difficulty) that may impact how many responses must be collected during item calibration. Practical and specific

guidelines regarding when sufficient information has been collected during Classical Test Theory item calibration are not available.

Due to the lack of specific and practical item calibration guidelines, the burden of item calibration is unlikely to be appropriate for calibrating items to enable accurate classification decisions. Instead, item calibration is likely too heavy or too light. Faced with uncertainty about how much information to collect, test administrators and researchers (Rudner, 2002a; 2009) understandably take the safe approach of basing item parameter estimates on large calibration sample sizes. However, the safe approach, involving hundreds or thousands of examinees, is not feasible in most educational contexts. As such, adoption of VL-CCT and availability of quality information about learner knowledge remains sparse.

In addition, there are several issues with the current state of VL-CCT research. First, studies involving the simulation of VL-CCT methods (Rudner, 2002a; 2009) use item information gathered during a simulated calibration phase involving a factor of ten more examinees than the number of examinees suggested by Frick (1992), casting into doubt the applicability of study findings to many real-world settings where a calibration sample of such size is impractical. Second, few studies (e.g. Tao et al., 2008) have been found that report on VL-CCT methods that have been used in real-world settings since the handful of studies conducted in the nineties (Welch & Frick, 1993; Welch, 1997) and these studies have limitations, recognized by the authors, including: the lack of a performance incentives for test takers, relatively small *N*'s impacting agreement calculations, clustering of scores around the cut-off, and possible violation of the assumption that the test measured a single unidimensional learning objective.

While VL-CCT approaches represent a promising approach for addressing the general problem of high resource requirements limiting application of CAT in instructional contexts, a specific problem must be addressed before the promise of VL-CCT can be realized – When has sufficient item calibration data been gathered to enable efficient and accurate VL-CCT?

1.3 Purpose

This research developed and evaluated a process for determining when sufficient item calibration data has been gathered to enable efficient and accurate VL-CCT. This dissertation presents a new VL-CCT calibration approach and an associated modification of the EXSPRT algorithm. The new VL-CCT calibration approach is labeled *Automatic Racing Calibration Heuristics* (ARCH). The modification of the EXSPRT algorithm involves the *measured* use of item-level parameter estimates only once they have become sufficiently calibrated.

The advantage of the ARCH approach is that with no or a very small initial item calibration phase CAT can begin and, through the course of testing, increasingly efficient VL-CCT approaches may be deployed that typically require a larger more laborious item calibration phase. ARCH places two VL-CCT approaches in a *race* to classify an examinee in as few items as possible. ARCH is *automatic* because item parameter estimates are continuously updated through the course of live testing after a small initial item calibration phase. Initially, availability of limited item calibration data would allow the SPRT VL-CCT approach to win the classification race because it would be the only approach with the data available to make a classification decision. However, the collection of additional item calibration data via live testing leads to increasingly precise item-level parameter estimates and makes the more efficient VL-CCT approaches that use item-level parameter estimates

and require more calibration data (i.e. EXSPRT) more competitive in the examinee classification race.

Specifically, this study:

(1) Examines how specific item characteristic measures can be leveraged to decide when an item-level parameter estimate has been sufficiently calibrated for use with EXSPRT without exceeding *a priori* established classification error rates,

(2) Compares the decisions reached by a joint application of SPRT and EXSPRT calibrated via ARCH to decisions reached by the total test and traditionally calibrated SPRT and EXSPRT,

(3) Compares the efficiency of a joint application of SPRT and EXSPRT calibrated via ARCH to traditionally calibrated SPRT and EXSPRT, and

(4) Tests the ARCH approach in both simulated and real-world testing environments.

1.4 Research Questions

1. When is an item sufficiently calibrated for use with the EXSPRT based component of ARCH? It was expected that item characteristic measures based on hypothesis testing techniques would indicate the discrimination and precision of item parameter estimates. Furthermore, it was expected that threshold values of these measures would indicate that sufficient information has been collected for an item such that use of the item parameters with EXSPRT would be unlikely to contribute to increasing classification error rates above those established *a priori*. However, since no previous studies have examined the use of hypothesis testing techniques to address item calibration issues in VL-CCT it could not be predicted which specific item

characteristic measures or values of item characteristic measures would best indicate that sufficient calibration data has been collected.

2. How accurate is ARCH in comparison to *a priori* error rates, SPRT, and EXSPRT? It was expected that ARCH, like traditionally calibrated SPRT and EXSPRT, would make classification decisions within error rates established *a priori*.
3. How efficient is ARCH in comparison to SPRT and EXSPRT? It was expected that before ARCH was using item-level parameters for all the items in the item-bank the proposed approach would be more efficient than SPRT but less efficient than EXSPRT. However, once ARCH uses item-level parameters for all the items in the item-bank, it was expected that the proposed approach would be more efficient than SPRT and as efficient as EXSPRT.

1.5 Significance

In general, this research helps to reduce barriers associated with item calibration that hinder adoption of VL-CCT methods in educational contexts in order to increase the quality of classroom assessment practice and provide educators and researchers with better and more timely information about learner knowledge. The potential of computers to enable more efficient classroom assessment has long been recognized. For example, with respect to using computers in classroom assessment Ferguson states “During the course of a school year, large numbers of hours now spent in testing could be invested in instructional activities or in supplementary diagnostic testing” (Ferguson, 1969, p. 15). Despite the recognition of the potential of computers to positively impact classroom assessment and the exponential advances in computing power since Ferguson’s statement, their use remains rare.

The significance of this study can be further explained by highlighting specific contributions it can make to: (1) the development of information-age learning management systems (LMS); (2) the evaluation of the effectiveness of instructional interventions particularly those involving classroom technology integration; (3) the establishment of a bridge (or arch) between national and international large-scale assessment and classroom assessment practice.

VL-CCT is a technology well suited to contribute valuable functionality to major roles in an information-age LMS. According to Reigeluth and colleagues (2008) the four primary roles for an information-age LMS are: record keeping, planning, instruction, and assessment. On first glance it may seem that VL-CCT only relates to assessment, however, the authors make it clear that the four roles are interconnected in an information-age LMS. Assessment produces information about learner knowledge that is the focus of record keeping and necessary for making evidence based decisions related to planning and providing learner-centered instruction. In turn, assessment data from the implementation of a personal learning plan (Watson & Reigeluth, 2008) and associated instruction are fed back into the LMS's record keeping system. Key decisions an information-age LMS are meant to support include determining if attainment of a specific standard is within the reach of a particular learner and if the standard has been met. Both are examples of classification decisions that VL-CCT can serve to inform. A LMS not able to promote efficient, valid, and reliable classification decisions about learner knowledge may not meet the definition of a truly information-age LMS given how critical revealing the state of learner knowledge before, during, and after instruction is to providing learner-centered instruction.

Computerized classification tests may be particularly appropriate for evaluating the effectiveness of classroom technology integration. The effectiveness of classroom technology integration can be measured in terms of the impact it has on promoting student learning (Brush, Glazewski, & Hew, 2008). However, integrating technology into classrooms pose challenges for the assessment of learning. If students are gaining knowledge and skills that take place in technology-enhanced environments (e.g. use of spreadsheet software to analyze data or evaluating if an online source is trustworthy) then assessment of student ability to perform technology-supported tasks may be difficult or impossible with traditional paper and pencil assessments. VL-CCT may be especially useful in assessing abilities that include a technological component for two reasons. First, test items, particularly those that include the use of innovative items (Parshall & Harmes, 2007), can be presented in a way that closely mimic or are indistinguishable from the authentic task to be accomplished with the aid of technology. Second, in VL-CCT examinee responses are provided in digital form and, as such, may be able to be automatically assessed without human involvement thus increasing efficiency and reducing the grading burden for educators. Furthermore, the need to assess abilities that include a technological component does not apply to K-12 students alone. Existing survey measures of teacher educational technology beliefs (e.g. Brush, 2008), technological pedagogical knowledge (e.g. Barrett, 2010), and technological pedagogical and content knowledge (e.g. Schmidt et al., 2009) demonstrate the considerable interest researchers have in determining the technology related skills of teachers as well as students.

VL-CCT is uniquely positioned to play a key role in balanced assessment frameworks designed to systematically serve complimentary but unique assessment functions. In a balanced assessment framework different levels of assessments align with each other in terms

of their learning objectives but their forms differ to meet competing formative and summative functions (Hickey et al., 2006). Hickey states that close level assessments focus on promoting student learning through the generation of formative information about a students' authentic practice often via discourse. Examples of discourse-based assessment include Duschl and Gitomer's assessment conversations (1997) and online formative peer assessment activities (e.g. Barrett & Howard, 2010). At the other end of the spectrum are distal assessments focused on the generation of standardized summative information (e.g. international large-scale assessment) that are primarily used to compare academic achievement among groups and typically apply IRT approaches to precisely place learner ability on a continuum (Hickey et al., 2006). Between distal and close assessments are proximal assessments that provide a blend of formative and summative information.

The three central ideas of the balanced assessment framework are: (1) different levels of assessment are needed to serve different functions and meet different stakeholder needs; (2) performance at one assessment level *should* reflect performance at other levels of assessment; and (3) if performance at one level is not reflective of performance at other levels then assessments in the framework are not sufficiently aligned and need adjustment (Hickey et al., 2006). The goal of an aligned balanced assessment framework is to enable educators to focus more on activities that promote student learning and less on test preparation while still positively impacting student performance on large-scale standardized tests.

VL-CCT have a unique and important role to play in balanced assessment frameworks. Classification tests are like norm-referenced large-scale standardized assessments in that they provide summative information about learner ability. However,

classification tests provide specific information about learner ability better suited for serving formative assessment functions aimed at promoting learning – specific feedback can be given to learners about what they know and aid selection of learning experiences to address particular gaps. For example, a large-scale standardized test may suggest that a student is in the seventieth percentile of fourth graders in terms of mathematical reasoning but what is an educator, parent, or student to do with such information if they want to improve the student's mathematical reasoning ability? On the other hand, information from a VL-CCT that indicates that the same student has mastered word problems that required single digit multiplication but has not mastered those involving single digit division helps to focus where the student does or does not require additional instruction and points the way to improving subsequent large-scale standardized test scores.

Finally, VL-CCTs could be extremely valuable in massively open online contexts (e.g MOOCs) that have been criticized as frequently providing "very little timely and informative feedback on learner performance" (Spector, 2014, p. 389). Efficient and accurate summative assessment is a key strength of VL-CCTs. Furthermore, limiting the number of items on a test exposes fewer items in the item bank to a given examinee, which serves to improve test security. If credentials awarded via MOOCs have value, some individuals will invariably attempt to cheat to earn the credentials. Use of VL-CCTs, particularly methods that use random item selection, can hinder cheating by making answer keys difficult develop through nefarious methods.

1.6 Definition of Key Terms Used

The following briefly defines key terms used and directs the reader to where additional information can be found. Here, a bottom-up approach is used to present the key

terms associated with items, tests, and finally key theories related to assessment to provide the reader with definitions of central assessment terms and concepts that are frequently referenced in the rest of the study.

1.6.1 Item Terms

Item. An item is the fundamental unit of a test and contains two distinct components: the stem and alternative of response actions (Burton et al., 1991). Traditional multiple-choice items are typically textual and contain a stem that poses a problem and alternative response actions (e.g. true/false or a set of alternatives labeled for easy identification). In the case of dichotomous scoring of items, which are evaluated as either correct or incorrect, there is a unique correct response action or answer with the remaining representing distractors. According to Burton (1991, p. 3) "The purpose of the distractors is to appear as plausible solutions to the problem for those students who have not achieved the objective being measured by the test item." Partial-credit or polytomous scoring, on the other hand, enables each alternative response action to be evaluated at a more granular level with a measure of the degree of correctness replacing a binary correct/incorrect evaluation.

Items, particularly when delivered via computers, need not be limited to text. However, a key restriction is that after presenting the item and allowing the examinee to respond that the response be sufficiently captured to enable, in the case of computerized testing, the response to be evaluated immediately (Parshall & Harmes, 2007). This restriction leaves room for a wide range of innovative item types that are discussed in detail by Parshall and Harmes (2007).

Item Parameter Estimates. Item parameter estimates are a set of measures that represent the probability of an examinee with specific characteristics providing a particular

response to an item. Item parameter estimates may take the form of a bounded set of probabilities or may be represented by a mathematical function that provides the probability of a correct answer associated with a continuous range of examinee ability values. In both cases, the item parameter estimates are typically established during item calibration and are used during testing to estimate an examinee's ability (either in the form of a classification decision or a point estimate of ability).

Item-bank. An item-bank is a set of items that can be selected for administration to an examinee during a test. Item-banks can be unidimensional or multidimensional. A unidimensional item-bank only includes items that measure the same construct whereas a multidimensional item-bank includes items that measure two or more constructs. Items may be selected from an item-bank randomly or using well-defined algorithms that use a variety of factors (e.g. current examinee ability estimate, item exposure rate, length of test, etc.) to decide which item to select next.

1.6.2 Test Terms

Test. The term test refers to a set of items that are presented to an examinee for the purpose of achieving a goal related to assessing examinee knowledge. A test may be fixed or variable in length. A test can always present the same items in the same order or present different items in different sequence. Tests may be delivered via computing resources, on paper, observation and assessment of a person's performance in a simulated or real context that requires application of knowledge or skills, or through some other method of communication (e.g. an oral test).

Test Administration. The term test administration refers to giving the test to a single examinee. While test length may refer either to the number of items presented in a specific

test administration or the length of time a specific test administration took, the prior is more common and will be the definition used in this text. Test duration refers to the time taken to complete a particular test administration. A test can be norm or criterion-referenced.

Norm-referenced Testing. Norm-referenced testing is when a test precisely estimates an examinee's ability typically for ranking or sorting purposes.

Criterion-referenced Testing. Criterion-referenced testing, on the other hand, places an examinee into one of two or more mutually exclusive groups (e.g. master or nonmaster; basic, proficient, or advanced ability, etc.) based on an "absolute standard of quality" (Glaser, 1963, p. 519).

1.6.3 Testing Theory Terms

Classical Test Theory and Item Response Theory (IRT) are two central theoretical approaches to testing.

Classical Test Theory. A key tenet of Classical Test Theory is that an examinee's score on a test (e.g. 85 correct out of 100) is based on two components: their true score and test measurement error. Test measurement error includes all the aspects of the test that may increase or decrease an examinee's score that have nothing to do with the examinee's ability with respect to the constructs that the test is supposed to measure. Cheating, guessing the correct answer, and poor item construction are examples of factors that increase test error. The true score is the score an examinee would get if measurement error had been entirely eliminated.

Item Response Theory. IRT, also known as latent trait theory, was developed by Frederic Lord and is the focus of the bulk of current testing research. Central to IRT is the

idea that the probability of an examinee with a particular ability providing a specific response can be modeled with a mathematical function. IRT can be applied in both criterion and norm-referenced testing context but requires that items be calibrated with a calibration sample that traditionally involves hundreds or thousands of examinees.

CHAPTER II. LITERATURE REVIEW

2.1 Overview

The literature review will first define Variable-Length Computerized Classification Testing (VL-CCT) and outline the specific goals and focus of the literature review. VL-CCT will then be situated within the larger context of approaches to testing and decision-making before getting into the specific design components of VL-CCT, which are used to narrow the scope of the literature review. Early research relevant to VL-CCT based on Classical Test Theory will be reviewed before delving more deeply into recent relevant research.

2.1.1 Definition of VL-CCT

Computerized Classification Testing (CCT) are tests that use computing technology to place examinees into two or more mutually exclusive groups (Spray & Reckase, 1996). Thompson (2007) distinguishes CCT approaches as being fixed-length (FL) or variable-length (VL) and suggests that the term computerized classification testing be reserved for "the broader topic of classification exams administered by computer" (Thompson, 2007, p. 1). VL-CCT uses information about examinees, their responses to items, and the test items themselves to administer items until specific termination criteria related to making a classification decision are met thus, unlike FL-CCT, allowing test length to vary. The goal of VL-CCT, the classification of an examinee into one or more mutually exclusive groups, is different from other testing goals such as precise estimation of an examinee's ability or determining an examinee's rank within a larger population.

2.1.2 Goals and Focus

The general goal of this literature review is to identify studies that examine VL-CCT approaches based on Classical Test Theory, subsequently referred to as classical VL-CCT.

Specifically, the literature review is focused on supporting two assertions that provide the impetus for the current study: (1) Classical VL-CCT approaches are viable for use in instructional contexts; (2) Lack of specific guidelines regarding the calibration of items for use with classical VL-CCT approaches hinder their application and adoption in instructional contexts.

A top-down approach is taken to first outline where VL-CCT fits within the larger context of testing and decision-making before presenting the five design components of VL-CCT in order to contrast classical VL-CCT with other approaches. The literature reviewed falls into two broad categories: historical studies that detail the development of classical VL-CCT and recent research that supports the viability of classical VL-CCT or helps to define the gap in our understanding of item calibration in classical VL-CCT.

2.2 Approaches to Testing and Decision Making

Understanding where VL-CCT based on the Classical Test Theory psychometric model fits into the larger context of testing requires that the associated nomenclature used be crystal clear. The lack of a standardized nomenclature in testing literature has been recognized (Thompson, 2007) and the use of acronyms that have common letters is frequent (e.g. CAT, CBT, and CCT), consequently, distinguishing among different types of test approaches can be challenging. Table 1 presents a proposed extended version of the nomenclature suggested by Thompson (2007) whose purpose is twofold: (1) to clarify test nomenclature for use in the remainder of the proposal and (2) to provide a framework for understanding what falls under the scope of the literature review, VL-CCT, and what does not.

2.2.1 Testing Attributes, Options, and Abbreviations

Commonly used terms such as computer adaptive testing (CAT) and computer based testing (CBT) have been criticized (Thompson, 2007; Rudner, 2009) for their lack the precision which has contributed to difficulties in distinguishing amongst different types of testing approaches and associated research. Every test has specific testing attributes with respect to length, deployment method, and goal. Each of these three attributes will be discussed in turn below.

Table 1. *Testing Attributes, Options, and Abbreviations Framework*

Testing Attribute	Available Options	Abbreviation
Length	Fixed-Length	FL
	Variable-Length	VL
Deployment Method	Computerized	C
	Traditional	T
Goal	Ability Classification	C
	Ability Estimation	E

The length of a test can be fixed or variable. A variable-length test, frequently referred to as an adaptive test, is any test whose length varies according to a pre-established set of rules that typically define the conditions under which the test terminates (e.g. confidence in a classification decision or precision of an ability estimate) (Thompson, 2007). The goal of variable-length testing is to achieve the purpose of the test (e.g. a classification decision or point estimate of ability) more efficiently than fixed-length tests without compromising reliability or validity. With fixed-length tests all examinees receive tests with the same number of items.

Table 2. *Example Tests Associated with Combinations of Testing Attributes*

	Length			
	Fixed		Variable	
	Deployment Method		Deployment Method	
Goal	Computerized	Traditional	Computerized	Traditional
Ability Classification	Current test associated with Indiana University plagiarism tutorial	Connecticut Mastery Test	National Council Licensure Examination – Registered Nurses (NCLEX-RN)	Many job interviews
Ability Estimation	Graduate Record Examination (GRE)	Trends in International Mathematics and Science Study (TIMSS)	Graduate Record Examination (GRE)	Binet IQ test (Binet & Simon, 1905)

There is an issue with focusing on test-length instead of whether the test is adaptive or not – it is possible for a fixed-length test to be adaptive. For example, a test that selects items from an item-bank based on previous examinee responses but always presents the same number of items would be considered both fixed-length and adaptive.

Fixed branching approaches to adaptive testing (e.g. Linn et al., 1969) provide tests where every examinee responds to the same number of items but selection of items is dependent on examinee responses and the location of items in a pyramid or tree structure. Given that adapting a test is typically associated with increasing test efficiency (i.e. enabling tests to terminate once specific conditions are met) the negatives associated with the increased complexity of adding an additional testing attribute to the framework in Table 1

(e.g. adding an *item selection* attribute with two options: *flexible* and *inflexible*) is viewed as greater than possible benefits that such an addition would provide.

Two primary methods for deploying a test are via computer resources or using traditional approaches (e.g. orally or using paper and pencil). Thompson (2007), in his effort to clarify nomenclature related to testing, suggests that variable-length testing requires the use of computing resources (Thompson, 2007). However, Thompson's suggestion could confuse rather than clarify since examples of variable-lengths tests that have not used computers to vary test length do exist. Early research on variable-length testing by both Binet (1905) and Hutt (1947) deployed variable-length tests where a human examiner took the place of the computer in selection of items to present to the examinee. The complexity of providing a variable-length test using traditional methods may make non-computerized deployment seem unreasonable. However, many job interviews and live performance tests are both variable in length and adaptive with on-the-spot human judgment applied to make decisions about what task or question the examinee should do next and when the test should end. Table 1 treats length and deployment method as two unique test attributes to avoid confusion and to leave room for the possibility of a variable-length test deployed via traditional methods.

The two primary goals of testing are classification or estimation of examinee ability (Rudner, 2009). Recall, classification decisions place an examinee into one of two or more mutually exclusive groups (e.g. master or nonmaster; basic, proficient, or advanced ability, etc.) based on an "absolute standard of quality" (Glaser, 1963, p. 519). A point estimate of ability, on the other hand, is based "upon a relative standard" (Glaser, 1963, p. 519) and

typically takes the form of a numerical score on a continuous scale (e.g. a score of 600 on the verbal component of the GRE).

The value of the framework presented in Table 1 can be illustrated by how it can be used clear-up confusion over the use of the term Computer Adaptive Testing. Welch (1997) asserts "Since adaptive tests are usually mastery-type tests, they are criterion-referenced as opposed to norm-referenced" (p. 9). Parshall and colleges (2002), on the other hand, contrast computerized adaptive testing (CAT) with computerized classification testing (CCT) by indicating that the former is focused on determining a point estimate of ability whereas the later classifies examinee ability into two or more categories. Thompson (2007) supports Parshall's perspective when he associates CAT with point estimates of ability that are typically applied in norm-referenced testing.

Using the term computerized adaptive testing to refer to tests whose goal is a point estimate of examinee ability is common practice (e.g. Chang & Lu, 2010) but may cause confusion since no mention is made in the term to estimation of ability. A computerized adaptive test with classification of ability as the goal can be reasonably viewed as a CAT. In contrast, Table 1 abbreviations clearly delineates between tests with different goals: VL-CCT approaches have the goal of classification and can easily be distinguished from VL-CET approaches whose goal is estimation of ability.

2.2.2 VL-CCT Focus

The focus of the literature review and the subsequent two studies are on variable-length computerized classification testing (VL-CCT) approaches because they enable efficient classification decisions about learner knowledge that educators frequently must make. A focus on VL-CCT removes all but one of the eight combinations of length, delivery

methods, and goal options presented in Table 2 from the scope of this review. However, design choices amongst various components for the construction of tests that apply VL-CCT approaches provide an additional opportunity for narrowing the focus particularly with respect to the psychometric model and item bank.

2.3 Design Components of VL-CCT

Thompson (2007) outlines five required design components (Table 3) that have much in common with components of computerized adaptive testing (Weiss & Kingsbury, 1984). Each of these five components is addressed below. Three of the components, psychometric model, item bank, and termination criteria, will receive the bulk of the attention given their important role in distinguishing VL-CCT based on Classical Test Theory or Item Response Theory (IRT) from other approaches.

Table 3. *Design Components of VL-CCT (adapted from Thompson et al., 2007)*

Design Component	Available Options	Example References
Psychometric Model	Classical Test Theory	Linn, Rock, & Cleary, 1972; Frick, 1992; Rudner, 2002
	Item Response Theory	Reckase, 1983; Kingsbury & Weiss, 1983; Lau & Wang, 1998; Eggen & Straetmans, 2000
Item Bank (kurtosis not always reported)	Peaked	Xiao, 1999
	Not Peaked	Kingsbury & Weiss, 1983; Finkelman, 2003; Yang, Poggio, & Glasnapp, 2006
Starting Point	Default (PR = 1 or $\theta = 0.0$)	Most extant research
	Previous information	Yang, Poggio, & Glasnapp, 2006
Item Selection	Random	Frick, 1989
	Estimate-based	Reckase, 1983; Kingsbury & Weiss, 1983; Eggen, 1999
	Cutscore-based	Spray & Reckase, 1994, 1996; Eggen, 1999
	Global (Mutual)	Weissman, 2004
Termination Criterion	SPRT	Reckase, 1983; Frick, 1989; Frick, 1992; Welch & Frick, 1993; Eggen, 1999; Eggen & Straetmans, 2000
	IRT Confidence Interval	Kingsbury & Weiss, 1983; Eggen & Straetmans, 2000; Change, 2006
	Bayesian decision theory	Vos, 2000; Glas & Vos, 2006

2.3.1 Psychometric Model

Classical Test Theory and IRT are the two dominate psychometric models that undergird testing research and practice. While both can be applied in VL-CCT and require a calibration phase to obtain item parameter estimates, each comes with unique constraints that must be considered when developing a VL-CCT for a specific context (Thompson, 2007). According to Thompson (2007, p. 3), "The first step in the technical development of a VL-CCT is the selection of a psychometric model that will be used as a basis for the remaining components."

The key idea of Classical Test Theory is that an observed test score is comprised of the examinee's true score and error (Novick, 1966). The true score is an error-free measurement of the desired examinee trait that cannot be directly observed. Error is the portion of the observed test score that can be contributed to factors not related to the trait being measured (e.g. guessing correct answers or poorly constructed test items). The relationship between the observed score, the true score, and error form the basis for examining the quality, specifically the reliability, of tests (Allan & Yen, 2002).

Item level statistics are also a key component of Classical Test Theory. The proportion of correct responses from examinees belonging to specific classification groups are used to establish difficulty and discrimination estimates for items and, less commonly, entire item-banks. For example, responses from known nonmasters to a specific item or a set of items making up a traditional test can be used to calculate how difficult the item or test is for nonmasters. Known masters would likely respond correctly more frequently to the same item(s) and, consequently, would have a unique difficulty estimate. The item-level application of Classical Test Theory form the basis for two related VL-CCT approaches:

Frick's modification of the sequential probability ratio test using expert systems reasoning labeled EXSPRT (1992) and Rudner's Measurement Decision Theory (2002).

The group (e.g. masters versus nonmasters) dependent nature of item difficulty and discrimination estimates has been identified as a key drawback of Classical Test Theory (Weiss & Yoes, 1991; Jacobs-Cassuto, 2005). With Classical Test Theory it cannot be assumed that the difficulty estimate of an item for any two groups will be the same. Furthermore, classical VL-CCT (VL-CCT that uses Classical Test Theory as the psychometric model) depends on the ability to able to distinguish between classification groups, ideally via a method independent of the specific test (e.g. a separate test, expert judgment, etc.) (Frick, 1992; Rudner, 2002; Thompson, 2007). However, distinguishing between classification groups may or may not be feasible in a particular testing context. Even with a clear method for distinguishing groups, two unique samples drawn from the same group are unlikely to result in identical item difficulty and discrimination estimates.

Fan (1998) explains that the possibility for a circular dependency is frequently cited as a major weakness of Classical Test Theory: "(a) The person statistic (i.e., observed score) is (item) sample dependent, and (b) the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample dependent." (p. 1). However, this circular dependency may be broken if items statistics are not (examinee) sample dependent. For example, Frick (1989) set item-bank level statics for use with SPRT without the use of sampling of examinees and instead applied established A through F grade cutoffs to define nonmastery and mastery groups.

IRT based methods aim to establish a precise measure of examinee ability that is then used to place an examinee into a classification group. Unlike Classical Test Theory, IRT

purports not to require an independent method for distinguishing between classification groups and it has been argued that it is not group dependent. Fan (1998) explains that, theoretically, IRT models are said to generate item statistics that are independent of examinee samples and this invariance property of IRT is the cornerstone of arguments that favor IRT over Classical Test Theory and used this to justify the complexity of IRT models. However, empirical investigation of the invariance property of IRT and Classical Test Theory based item and examinee statistics has brought into question the validity of the invariance property advantage of IRT over Classical Test Theory (Fan, 1998; Macdonald & Paunonen, 2002; Xu & Stone, 2012). Beyond the questionability of the invariance property, IRT comes with it's own set of drawbacks, restrictions, and assumptions that impact it's viability for use with VL-CCT.

According to Thompson (2007), the main drawback of using IRT in VL-CCT is the large calibration sample required for establishing item parameters – often ten times more examinees than those required for calibration with classical VL-CCT. Details of the item-bank calibration requirements for IRT will be discussed in the next section on the item-bank. IRT has more restrictive assumptions than does Classical Test Theory. IRT is based on three assumptions: (1) a unidimensional construct is being measured; (2) the local independence of items; (3) that the item response function selected reasonably models how examinees actually respond to items. Choice of the model used for the item response function may add additional constraints. In comparison, Classical Test Theory is based on only the assumption of the independence of items (Rudner, 2002), which makes ensuring that underlying assumptions of the psychometric are not violated an easier task.

The less restrictive assumptions associated with Classical Test Theory has been identified as a key advantage over IRT. In regard to measurement decision theory that is based on Classical Test Theory Rudner (2009) states:

> Thus, the tested domain does not need to be unidimensional, examinee ability does not need to be normally distributed, and one doesn't need to be as concerned, with the fit of the data to a theoretical model as is the case with IRT or in most latent class models.
>
> Rudner, 2009, p. 1

In addition, the validity of approaches for estimating item discrimination and lower asymptotes used with the two- and three-parameter IRT models have been questioned by advocates of the one-parameter IRT model (Wright, 1977).

2.3.2 Item-Bank

The item-bank is the pool from which the VL-CCT draws items. Important characteristics of an item-bank include size, the type of information known about items in the bank, and kurtosis of either individual items or the entire item-bank. The size of an item-bank is simply the number of individual items that it contains. Factors that impact item-bank size include the stakes involved in the test and the underlying psychometric model with higher stakes tests requiring more items and IRT enabling fewer items to be used if the items are highly informative (Thompson, 2007). Weiss (1985) suggests that an item bank in the range of 150 to 200 items typically provides reasonable results.

Both the type of information needed about items in the item-bank and the associated calibration requirements depend on the underlying psychometric model. Use of Classical

Test Theory will require either test-level or item-level parameter estimates be established. However, the bulk of the literature on item calibration requirements focuses on IRT. In fact, when outlining the major steps for calibration of an item bank Eggen (2007, p. 7) assumes IRT as the psychometric model and makes no reference to Classical Test Theory.

During the calibration phase for IRT data is collected about each test item that is then used to develop a theoretical model, the item response function, for the probability of a correct response to the item from an examinee with a given ability level. The choice of the IRT model (1PL-, 2PL-, or 3PL-) has implications on item calibration in terms of the sample size needed and to what extent the data must fit the model (Eggen, 2007) – larger sample sizes are required for models with more parameters. Simpler IRT models require hundreds of examinees in the calibration sample and more complex IRT models require a calibration sample of thousands (Wainer & Mislevy, 2000; Weiss & Kingsbury, 1984; Welch & Frick, 1993).

Kurtosis is an attribute of an item or test when IRT is the psychometric model. Kurtosis refers to the examinee ability level at which point an item or a set of items provides the most information about the examinee. Kurtosis is used in intelligent item selection algorithms (discussed further below), such as the maximum information search and selection method (Kingsbury & Weiss, 1983), to select items that will provide the most data about an examinee given the current estimate of their ability (Rudner, 2002).

2.3.3 Starting Point

The starting is the best guess regarding the probability of an examinee belonging to a specific classification group before the test has begun. In the absence of information, the examinee is assumed to have an equal probability of belonging to each of the possible

classification groups. However, if information is available (e.g. the score an examinee has obtained on a previous related test), then this information may be used to modify the starting point (Thompson, 2007; Weiss & Kingsbury, 1984; Yang, Poggio, & Glasnapp, 2006).

2.3.4 Item Selection

Items can be either selected from the item-bank randomly or intelligently. With random item selection all items in the item-bank have an equal chance of getting selected for administration next. Intelligent-item selection purposely selects items based on what is known about the examinee and the items in the item-bank. Detailing the plethora of intelligent item selection algorithms available for use with VL-CCT is beyond the scope of this review. Interested readers are directed to reviews of intelligent item selection procedures (Thompson, 2007; Rudner, 2009).

While many of the item selection procedures are based on using IRT as the psychometric model, use of Classical Test Theory does not preclude using information about items to intelligently select items for presentation to the examinee. For example, EXSPRT-I (Plew, 1989; Frick 1992) uses Classical Test Theory item parameter estimates to intelligently select items "that best discriminates between masters and nonmasters and is least incompatible with the current estimate of the examinee's achievement level" (Welch, 1997). Rudner (2009) details three intelligent item selection methods for use with Classical Test Theory: minimum expected cost, information gain, and maximum information.

2.3.5 Termination Criterion

Termination criteria define when a VL-CCT can make a classification decision about an examinee. According to Thompson (2007), three types of termination criteria are used with VL-CCT: IRT-based confidence intervals, a Bayesian decision theory framework

(Lewis & Sheehan, 1990; Sheehan & Lewis, 1992), and Wald's (1947) sequential probability ratio test (SPRT).

Termination criteria based on confidence intervals continue testing until confidence in the estimate of the examinee's ability reaches a point where the confidence interval lies completely within the range of ability associated with one of the classification groups (Eggen & Straetmans, 2000; Thompson, 2007). Since the confidence interval approach depends on a point estimate of examinee ability it is inappropriate for use with classical VL-CCT.

Application of a Bayesian decision theory framework for determining when a test can end enables the inclusion of costs structures associated with making specific classification mistakes and of administering an additional item (Thompson, 2007). For example, consider the costs of misclassifying two true nonmasters A and B as masters where A is very close to being a master and B is very far from being a master. The potential negative impact that B could cause as an incorrectly labeled a master in the real-world would likely be substantial greater than A and use of a Bayesian decision framework enables these potential costs to be factored into termination criteria (Vos, 1999).

With sequential probability ratio test (SPRT) choice among classification options is treated as a statistical hypothesis-testing problem (Eggen & Straetmans, 2000; Thompson, 2007). SPRT was originally developed by Wald (1947) for use in manufacturing to determine if the quality of a batch of goods (e.g. ammunition) warranted their rejection or not. The advantage of SPRT over other statistical approaches associated with fixed sample sizes is the efficiency of decision-making (Frick, 1989). Applying SPRT after each observation (e.g. is a given bomb in a shipment a dud?) enables decisions to be reached in a way that minimizes

the observations necessary, an important factor when there are costs associated with observing (e.g. destruction of a bomb or exposure of an item).

The SPRT originally only applied to making decisions between two alternative hypothesis. However, Armitage (1950) asserted that Wald's theory of sequential tests for deciding between two alternative simple hypotheses could be extended to deciding between several alternative simple hypotheses where decision error can be controlled. Using mathematical argument and inspection diagrams, Armitage demonstrates a procedure for making decision between *k* alternative hypotheses. An "arbitrary constant" A is used that seems to be related to type I or type II errors in choosing one hypothesis over the other but the nature of this relationship is not clearly explained.

2.4 Classical VL-CCT Research

2.4.1 Focus on Classical VL-CCT

The rest of the literature review and the subsequent studies are focused on classical VL-CCT since use of IRT as the psychometric model requires a calibration sample that is impractical most instructional contexts. In comparison to research on IRT based VL-CCT, there has been relatively little research focused on classical VL-CCT. Thompson (2007) provides three example references of classical VL-CCT studies: Linn, 1972; Frick, 1992; Rudner, 2002. Rudner (2009) identifies an additional three articles that illustrate the item-level application of what he refers to as measurement decision theory: Macready and Dayton, 1992; Vos, 1997; Welch and Frick, 1993.

The following will review a number of the foundational studies that focus on classical VL-CCT and summarize relevant findings. The argument that classical VL-CCT is a viable approach for use in instructional contexts is supported by research that shows that relatively

small calibration sample sizes enable classical VL-CCT approaches to efficiently make classification decisions within acceptable error rates. Early research relevant to classical VL-CCT focused on application of the SPRT using test-level item parameters either using traditional deployment methods or early computers. Use of item-level parameters occurred with Frick's later application of expert systems reasoning to the SPRT.

2.4.2 Early Research Relevant to Classical VL-CCT

The first reported application of the Wald's (1947) SPRT in the context of making classification decisions about examinees comes from Cowden (1946). Cowden's aim was to determine if the SPRT could be used to assign A through F grades on a test of elementary statistics. The test consisted of 100 true/false items that were administered to ten examinees in batches of 20 items. At the end of each batch the number of errors made by an examinee was compared to a table that outlined the termination criteria associated with passing and failing based on a pre-established sequential sampling plan.

When a pass/fail decision could be made the examinee was given the option of continuing with another batch or accepting the grade they had been assigned. The differences between the consequences of passing or failing on the ultimate grade assigned grew less pronounced as additional batches of 20 items were given. For example, passing after the first batch earned an A but failing resulted in an F. However, after four batches a pass earned a C and a fail still resulted in an F.

Results of Cowden's study demonstrated that SPRT could be used to shorten test length from 100 items to 63 items on average. However, it is unclear if the increased efficiency came at the cost of accuracy since the accuracy of the classification decisions was not evaluated. Also, the administration of items in batches in Cowden's study may violate the

assumption of the independence of items since examinees could have used information in one item to modify their responses to another item. The mapping between scores at various stages and grades imply that some type of test-level parameter estimates were being applied but the details of which are not explained. Finally, it seems that no attempt was made to identify items that could possibly be problematic (i.e. those that participants with higher grades found more difficult than did participants with lower grades).

A study by Moonan (1950) is noteworthy because it was the first to apply SPRT retroactively to examinee response data and simulate both random item selection and the application of SPRT after each item administration. The test used in Moonan's study consisted of 75 multiple-choice items each having five response options. 39 examinees took the complete test and their responses were then used in simulating five different SPRT tests that each had unique cutoff scores and error criteria. Correlations between the simulated SPRT tests and the total scores indicated a high degree of consistency in classification decisions. Furthermore, the average number of items administered before a SPRT decision could be made was around 40 representing a considerable efficiency gain. Again, no attempt was made to evaluate if items were problematic nor were test level item parameter estimates based on empirical data.

The first application what could truly be considered classical VL-CCT to improve test efficiency is from Ferguson (1969) since testing involved the use of computing resources. This research developed and field-tested a model for computer-assisted testing that uses an examinee's previous responses, entered via a "teletypewriter", to branch them to objectives and associated items that were "tailored to the competencies of the examinee" (Ferguson, 1969, p. 1).

Ferguson's study involved 75 elementary students in grades one to six that were placed into low, middle, and high proficiency groups by a coordinator. Participants had achieved different states of progress through associated math instruction (none to complete). All were given a computer adaptive test of basic math proficiency twice with no instruction in between. The math proficiency tested was based on the math problems that would typically be taught in grades three and four.

Each test focused on 18 objectives associated with addition and subtraction that were placed in a hierarchy based on hypothesized prerequisite dependencies between objectives. Several sequences of objectives were derived from the hierarchy that all ended in an objective that was not a prerequisite for any other objective. Items associated with each objective were generated on the fly by computer via selecting specific numbers to insert into the item stem and calculating the corresponding correct answer so that the examinee response could be evaluated.

All examinees started on testing with the same objective in the middle of the objective hierarchy. After an examinee response to an item associated with the objective, Wald's (1947) sequential probability ratio test was used to determine if the examinee responses warranted a classification decision regarding the objective. If a classification could be made regarding mastery of the objective, the next objective they would face would be lower in the particular objective sequence if they had not mastered the objective or higher in the sequence if they had mastered the objective. The process of generating items associated with the objective, evaluating mastery, and selecting a higher or lower level objective in the sequence would continue until one of two conditions were met: (1) the examinee did not master the current objective and no more untested objectives lower in the sequences remain;

(2) the examinee did master the current objective and no more untested objectives higher in the sequence remain. After completing the computer adaptive test examinees responded to a 150-item paper and pencil test that evaluated performance on all 18 objectives.

Ferguson found minimal inconsistencies in the results, that is, cases where objective A had not mastered when objective A is prerequisite to objective B and objective B had been mastered. Results also demonstrated that the computer adaptive testing approach had predictive validity: CAT performance was predictive of performance on the paper and pencil test. A high correlation was found between the two tests in terms of objectives mastered providing evidence of reliability and decision accuracy. The key finding was that the computer adaptive version of the test required nearly one-third the number of items as the full test and less than half the time. Also, fewer items were required to make nonmastery decision versus mastery decisions.

Several criticisms can be made about Ferguson's study. The exact procedure used to make routing decisions about examinees during testing and the criteria was used to place participants into the low, middle, and high proficiency groups lacked detail that would enable the study to be replicated. Furthermore, having 1st through 6th grade students take a test designed for 3rd and 4th graders raises questions regarding the appropriateness of the sample selected for the study. The choices of the probabilities of a correct response from a master or nonmaster and the error rates are not based on empirical data. Again, no attempt was made to identify problematic items and the test-level probability estimates are not justified in empirical data. However, Ferguson's research does provide an early example of the power of classical VL-CCT approaches to impact test efficiency while maintaining classification accuracy.

Linn (1972) used actual examinee response data to test theoretical assertions by Green (1970) that sequential testing could yield 50 percent shorter tests without compromising classification accuracy. Responses from nearly 5,000 examinees on General Examinations of the College Board's College Level Examination Program (CLEP) were split randomly into two groups: half for item level calibration and half for "cross-validation". A cut-score was established that divided the group into approximately half: a higher scoring group and a lower scoring group.

Two sequential procedures were developed based on work conducted by Armitage (1950) for applying Wald's sequential ratio probability test to classification decisions. The first sequential procedure treated dimensions (Math, English Composition, and Natural Sciences) separately and the other used Math score in estimating other scores. The experiment was repeated for increasingly lower probabilities of classification error and for conventional tests of twelve different lengths (5, 10, … , 55, 60). The assignment of examinees to classification groups by the sequential procedures was compared to assignments made using the results from conventional tests.

The findings supported the theoretical assertion that approximately half as many items are required with sequential testing to achieve same accuracy as conventional testing. Taking one dimension into account (e.g. Math) when testing another dimension was only useful when there is a strong correlation between the two (e.g. Math and Natural Sciences).

One issue with Linn's study was that cut-scores used were not based on achievement of an absolute standard. Instead the sample was just divided in two. Also, using over two thousand examinees for item calibration is not practical in many contexts and reasons were

not provided for setting sixty as the maximum number of items or the size of calibration sample size. It is not clear if any attempt was made to identify items that were problematic. A further criticism is that the types of possible error were not distinguished from each other but the author recognized this issue.

Kingsbury and Weiss (1983) compared the SPRT to IRT based approaches for making mastery decisions. Monte Carlo simulations were used to generate items, examinees, and examinee responses. Results showed SPRT was the most efficient approach but was also the least accurate. A major issue with this study is that incorrect equations for SPRT were used (Frick, 1990; Welch, 1997). Also, use of Monte Carlo simulations to generate data depends on an underlying model for the relationship between items, examinees, and examinee responses (e.g. examinee abilities theta values that are normally distributed) that may or may not reflect real world data.

2.4.3 DAL and COM Studies

A series of studies by Frick, Plew, and Welch represents a major step in classical VL-CCT: application of item-level parameter estimates to make additional efficiency gains. Research conducted by Frick (1989) found that the SPRT was a viable option for efficiently and accurately make mastery decisions despite variations in item difficulty and discrimination power and, consequently, could be leveraged to individualize learning experiences. Frick found that SPRT had high predictive validity.

Frick's (1989) research conducted computer simulations based on historical test data from two tests, a Digital Authoring Language (DAL) test and a test of knowledge of how computers functionally work (COM test). Both tests were delivered via computer where items were randomly selected from the item-bank without replacement until all items were

used. Also, both tests contained a variety of item types where difficulty and power of discrimination varied considerably.

The DAL test (97 items) was administered 53 times. Most DAL test examinees were graduate students taking a class with Frick that covered DAL programming. The remaining DAL test examinees were professional staff at Indiana University who claimed or did not claim to be experienced DAL programmers. Knowledge of the individual examinees previous experience with DAL enabled master and nonmaster groups to be defined independent of the examinee score on the DAL test. The DAL test (mean score 63%, *SD* = 24.6) was considerably harder than the COM test (mean score 79%, *SD* =13.6).

The COM test (85 items) was administered 104 times. Note that the original publication reporting on the COM test indicates that "There were 105 administrations of the COM test" (Frick, 1989, p. 102) but subsequent publications and the available historical data files show report the number to be 104 (Frick, 1992, p. 203). Current or former graduate students, representing two thirds of the COM test examinees, took the test twice at different points in a course and undergraduate students, representing the rest of the COM test examinees, took the test only once.

SPRT parameters for both tests were set to $P(C|M)$ = .85, $P(C|N)$ = .60, false master = false nonmaster = .025. Choice of these particular SPRT parameters were based on obtaining a grade of B or higher (> .85) or D or lower (<.60). Simulated SPRT tests were conducted post-hoc based on historical data from the DAL and COM tests. SPRT mastery decisions were compared to mastery decisions based on total test scores. Total test scores

were converted into mastery decisions using the mid-point between mastery and nonmastery (72.5%) as the cut score.

The 1989 study Frick found a high level of agreement between SPRT decisions and total score decisions: 96% agreement for DAL, 99% agreement for COM; 98% agreement across both tests. Fewer classification errors occurred than were expected (<5%). The mastery decisions were made via SPRT with mean test lengths that were less than one fourth of the total test length. On average, fewer items were required to make nonmastery decisions than mastery decisions. All SPRT classification errors were cases where it classified a master as a nonmaster.

In a follow-up study using test re-enactments with the same historical data from the COM and DAL tests Frick (1992) examined the efficiency and accuracy of several classical and an IRT based VL-CCT approaches where items were calibrated using different sample sizes. Frick introduces the EXSPRT-R, which uses item level parameter estimates and random item selection to make classification decisions about an examinee. Also introduced is the EXSPRT-I, which uses item level parameter estimates but applies intelligent item selection to make classification decisions. EXSPRT-I was jointly developed by Frick and Plew (1989) and applies item selection reasoning based on item discrimination, the item/examinee incompatibility, and the utility of the item.

While the focus of the study was on examining the accuracy and efficiency of the various VL-CCT approaches, an additional factor was also examined – the consequences of calibration sample size. For both the DAL and the COM test item parameter estimates were established using two different sample sizes: 25 and 50 randomly selected examinees with

the later included the former. The number of examinees who took the COM test also enabled calibration samples of 75 and 100 examinees to be used.

Results from the 1992 study showed that calibration sample size did not substantially impact test efficiency but did impact accuracy. When only 25 examinees were used to calibrate items both the accuracy of the decisions reached by EXSPRT-I and the EXSPRT significantly departed from decisions made using the total test based on a Chi-squared goodness of fit test ($p < .05$) and were less accurate than expected. With 50 examinees in the calibration sample all the approaches, except AMT and EXSPRT-I with the COM test, had classification accuracies within expected error rates that did not significantly differ from decisions made using the total test. Calibration samples of 75 and 100 examinees enabled all but the AMT approach to make classification decisions within *a priori* error rates.

Percent agreement numbers that were used in the study, unlike Proportion Reduction in Error (see Rudner, 2009, p. 7), do not explicitly address agreement due to chance. In addition, it is not clear if or how problematic items were handled. Welch (1997) points out that the SPRT tests in the Frick study were simulated rather than controlling the test in real-time which provided the motivation for subsequent studies by Welch and Frick.

Welch and Frick (1993) showed that SPRT and EXSPRT-R testing approaches can make accurate and efficient mastery decisions in real-time testing situations and are viable and practical alternatives to IRT based methods. Thirty-eight students from a graduate course on the use of computers in education were randomly assigned to two groups (20 given EXSPRT-R/SPRT and 18 given EXSPRT-I). Tests drew from an item-bank of 85 items that represented a variety of item types.

Examinees were told they would be taking two tests (adaptive and fixed length) but only one was truly given. Decisions were made at various points about examinee mastery using different algorithms but all examinees ended up taking all 85 items. Item parameters for SPRT, EXSPRT, and Rasch estimates were based on historical data from 185 administrations from past studies (Frick, 1989; Powell, 1992). For SPRT the probability of a correct response from a master was set at .90 and the probability of a correct response from a nonmaster was set at .63. Equal prior probabilities of master and nonmastery were assumed and the acceptable rate of false mastery and false nonmastery were both set to .01. The Adaptive Mastery Testing (AMT) method (Weiss and Kingsbury, 1984) was used for the IRT approach.

Results again showed that EXSPRT-I tests were significantly shorter than EXSPRT-R (half as long) but no significant differences were found among other tests. The conventional proportion correct with a confidence interval based on a standard error of measurement and the IRT theta estimation with a standard error of measurement based on test information at the given theta level made identical mastery decisions with both being unable to make decisions in nearly 40% of cases. EXSPRT-R procedures applied to the total test, on the other hand, made decisions in all but 13% of cases (one third of 40%). When compared to classification decisions made by applying EXSPRT-R procedures to the total test, EXSPRT-I disagreed in over 20% of cases and SPRT disagreed 10% of the time. Decisions made with AMT disagreed with EXSPRT-R procedures applied to the total test in over 20% of cases. SPRT performed about as well as other methods

On critique of the Welch and Frick (1993) study is that SPRT is presented as requiring no historical data to for probabilities of a correct response from a classification

group. However, this is not necessarily true. Decision makers can set these values but without empirical data to support their decision the accuracy of their estimates cannot be determined. Furthermore, it is not clear why "a conventional proportion correct metric with a .85 cut-off score" (Welch & Frick, 1993, p. 58) was used rather than the halfway point between the mastery and nonmaster SPRT probability of a correct answer as was done in earlier studies (e.g. Frick, 1989).

Only two studies could be found that specifically focus on how the size of the calibration sample impacts subsequent classical VL-CCT efficiency and accuracy – Frick's 1992 study already described and a study from Rudner (2009) that will be reviewed next.

2.4.4 Measurement Decision Theory

Rudner (2009) provides evidence that VL-CCT based on measurement decision theory (MDT) are as good or better than IRT-based approaches in terms of both classification accuracy and test length.

Rudner performed simulations using two simulated examinee datasets: randomly drawing an examinee ability level from N(0, 1) or U(-2.5, 2.5) and assigning individual to classification based on ability level and cut score for the particular test. Tests items parameters were based on historical data from 1999 Colorado State Assessment Program (CSAP) fifth-grade mathematics test (Colorado State Department of Education, 2000) and the 1996 National Assessment of Educational Progress (NAEP) State Eighth Grade Mathematics Assessment (Allen, Carlson, and Zelenak, 1999).

Classification accuracy was measured in two ways: proportion of correct state classifications and Proportion Reduction in Error (PRE). Test length was varied from a

maximum of 3 items to the size of item-bank. For each combination of conditions (test length, test, examinee data set) 1,000 administrations were simulated 100 times.

Rudner found that the PRE for MDT approaches were the same or higher than IRT approaches under all but one conditions (NAEP, U, max test length < 30). For item selection techniques "minimum cost and information gain decision theory approaches consistently out-performed the first two IRT approaches, and out-performed the IRT cut score approach when 20 or fewer items were administered" (Rudner, 2009)

It is unclear why the 41st percentile was used as cut score for the CSAP test. Also, it was not clear on how item parameter estimates were estimated beyond stating that "The latent states and the response vectors were used to compute the conditional prior probabilities of each response z_i given each mastery state m_k, $P(z_i|m_k)$." (Rudner, 2009, p. 7). It would have been helpful to have details such as how a response was simulated for a given examinee and a given item. Furthermore, the calibration sample sizes used were not sufficiently clear. Lastly, no attempt was made to identify problematic items and the use of thousands of examinees to set classical VL-CCT item-parameter estimates seems like overkill given Frick's (1992) demonstration that calibration sample sizes as small as twenty-five per classification group were sufficient.

2.5 Conclusion

The general goal of this literature review was to identify empirical studies that examine VL-CCT approaches based on Classical Test Theory. First, a top-down approach was taken to outline where VL-CCT fits within the larger context of testing and decision making before exploring the empirical research relevant to classical VL-CCT. Findings were

presented in the context of supporting two assertions that provides the impetus for the two studies conducted: (1) Classical VL-CCT approaches are viable for use in instructional contexts; (2) Lack of specific guidelines regarding the calibration of items for use with classical VL-CCT approaches hinder their application and adoption in instructional contexts.

CHAPTER III. EXPLICATING AUTOMATIC RACING CALIBRATION HEURISTICS

The Automatic Racing Calibration Heuristics (ARCH) approach is based on two existing VL-CCT approaches (SPRT and EXSPRT-R), a Bayesian statistics perspective on Classical Test Theory item calibration, and statistical approaches used in hypothesis testing. Each will be explained prior to introduction of the ARCH approach. The following will first illustrate, through examples, how SPRT and EXSPRT make classification decisions.

3.1 Classical Test Theory Item-Bank Calibration

There are multiple levels of Classical Test Theory calibration available with each level representing a tradeoff between ease of calibration and test efficiency. Consider the hierarchy behind a standard multiple-choice test. At the top of the hierarchy is the item-bank and at the bottom is a specific response to an item. The item-bank contains many items and each item has multiple response options.

Table 4 introduces a framework that reflects this hierarchy. At level one no probability calibration information has been gathered and instead decision-makers set cut-scores between classification groups using one or more approaches recommended for criterion-referenced testing (CRT) (Shrock & Coscarelli, 2007) such as the Angoff method (Angoff, 1984) or the contrasting groups approach. Tests are fixed length at level one since cut-scores are defined by a fixed number of correct responses out of a fixed number of total items and thus provide no opportunities for decreasing test length. Level 2 of classical calibration introduces probability estimates for responses for each classification group but does so at the item-bank level. Level 3 establishes item-level probabilities. Level 4 establishes response-level probabilities.

Table 4. *Four Levels of Item Bank Calibration with Classical Test Theory*

Calibration Level	Description	Example Approaches	Test Efficiency	Ease of Calibration
1. Cut-score	Cut-scores between classification groups established by decision-makers	Angoff method (Angoff, 1984); CRT (Shrock & Coscarelli, 2007)	Lowest	Highest
2. Item-bank	Probability of classification group answering randomly selected item correctly established	SPRT (Frick, 1989)	Low	High
3. Item (Dichotomous scoring)	Probability of classification group answering specific item correctly established	EXSPRT-R, EXPRT-I (Frick, 1992; Welch & Frick, 1993); MDT (Rudner, 2002; 2009)	High	Low
4. Item Response (Polytomous scoring)	Probability of classification group selecting a specific item response established	None found	Theoretically Highest	Theoretically Lowest

Higher-level classical test calibration typically results in higher test efficiency. Research has shown that VL-CCT methods based on calibration at level 2, such as the

sequential probability ratio test (SPRT) (Frick, 1989) and level 3, such as an extension of SPRT that applies expert systems reasoning with random item selection (EXSPRT), have compared favorably with traditional fixed-length tests (level one) in terms of reliability and validity while substantially reducing test length (Frick, 1992). However, VL-CCT methods that apply level 3 calibration have been shown to be even more efficient (shorter mean test length) than VL-CCT methods that apply level 2 (Frick, 1992; Welch & Frick, 1993).

Termination criterion used in level 3 VL-CCT methods leverage information about items and previous examinee responses to increase test efficiency. Some level 3 VL-CCT approaches, such as another extension of SPRT that applies expert systems reasoning and intelligent item selection (EXSPRT-I), use item information and examinee responses when selecting items to produce even more efficient tests (Frick, 1992). However, intelligent item selection can also make cheating easier if examinees can accurately predict which items they will face and memorize the relatively short sequence of correct answers necessary to produce the desired classification decision (Frick, 1992; Welch & Frick, 1993).

The increased efficiency of higher-level classical test calibration methods comes at the cost of increased effort dedicated to item calibration. In addition, the item calibration phase is made more difficult by the lack of clear guidelines on when calibration is sufficient to enable accurate classification decisions.

3.1.1 Cut-Scores

Cut-Scores can be established via expert opinion and or the gathering of empirical data from known masters and nonmasters. While there are a wide variety of techniques available for identifying one or more cut scores between classification groups none preclude the application of the judgment of test administrators (Shrock & Coscarelli, 2007). Of

particular concern are the consequences of making classification errors in establishing cut-scores.

3.1.2 Item-bank Level Probabilities

The expected mean of a beta variable (equation 1) can be used to estimate the expected future probability of success given a previous set of observations that can be classified as successes (*s*) or failures (*f*).

$$E[beta(* \mid s, f)] = \frac{s+1}{N+2} \quad (1)$$

Where:

*E[beta(* | s, f)]* = Expected mean of a beta variable with parameters *s* and *f*

s = Number of successes

N = Number of observations *(s + f)*

f = Number of failures

One application of the mean of a beta variable is to establish item-bank level probabilities. Equation 2 applies equation 1 in the context of establishing the probability of a master responding correctly to an item randomly drawn from the item-bank (*P(C|M)*). In equation 2 the number of correct responses by previous masters to item in the item-bank ($\#C_M$) takes the place of number of successes (*s*) and the number of incorrect responses by masters ($\#\neg C_M$) takes the place of number of failures (*f*). Since the only two outcomes for evaluating the response of a master is correct or incorrect the probability of a correct response from a master plus the probability of an incorrect response from a master must equal one (equation 3). In the same way the probability of a correct response (equation 4) and

an incorrect response (equation 5) from a nonmaster may be established. Only the four probabilities associated with equations 2, 3, 4, and 5 are established at the item-bank level.

$$P(C|M) = \frac{\#C_M + 1}{\#C_M + \#\neg C_M + 2} \quad (2)$$

$$P(\neg C|M) = 1 - P(C|M) \quad (3)$$

$$P(C|N) = \frac{\#C_N + 1}{\#C_N + \#\neg C_N + 2} \quad (4)$$

$$P(\neg C|N) = 1 - P(C|N) \quad (5)$$

Where:

$P(C|M)$ = Probability of a *correct* response given *mastery*

$P(\neg C|M)$ = Probability of an *incorrect* response given *mastery*

$\#C_M$ = Number *correct* responses from *masters* to items in the item-bank

$\#\neg C_M$ = Number *incorrect* responses from *masters* to items in the item-bank

$P(C|N)$ = Probability of a *correct* response given *nonmastery*

$P(\neg C|N)$ = Probability of an *incorrect* response given *nonmastery*

$\#C_N$ = Number *correct* responses from *nonmasters* to items in the item-bank

$\#\neg C_N$ = Number *incorrect* responses from *nonmasters* to items in the item-bank

3.1.3 Item-Level Probabilities

The establishment of item-level probabilities is very similar to establishing item-bank level probabilities with the key difference being that probabilities are established at the item level rather than the item-bank level. This results in four probabilities *per item in the item-bank*. Equations 6 through 9 are the item level equivalents to the item-bank level equations 2 through 5 with the only difference being the inclusion of the *i* subscript to represent that the equations apply to a single item *i*.

$$P(C_i|M) = \frac{\#C_{iM} + 1}{\#C_{iM} + \#\neg C_{iM} + 2} \quad (6)$$

$$P(\neg C_i|M) = 1 - P(C_i|M) \quad (7)$$

$$P(C_i|N) = \frac{\#C_{iN} + 1}{\#C_{iN} + \#\neg C_{iN} + 2} \quad (8)$$

$$P(\neg C_i|N) = 1 - P(C_i|N) \quad (9)$$

Where:

$P(C_i|M)$ = Probability of a *correct* response to item *i* given *mastery*

$P(\neg C_i|M)$ = Probability of an *incorrect* response to item *i* given *mastery*

$\#C_{iM}$ = Number *correct* responses from *masters* to item *i*

$\#\neg C_{iM}$ = Number *incorrect* responses from *masters* to item *i*

$P(C_i|N)$ = Probability of a *correct* response to item *i* given *nonmastery*

$P(\neg C_i|N)$ = Probability of an *incorrect* response to item *i* given *nonmastery*

$\#C_{iN}$ = Number *correct* responses from *nonmasters* to item *i*

$\#\neg C_{iN}$ = Number *incorrect* responses from *nonmasters* to item *i*

3.1.4 Item-Response Level Probabilities

Theoretically, the probability of a particular classification group selecting a specific item response could be established using the same equations described above. While a IRT approaches that apply polytomous scoring methods do predict the probability of specific item-responses being selected, they do so based on a continuous ability estimates rather than examinee classifications. However, such an item-response level application has not been found in the classification testing literature and will not be applied in this study.

Consequently, additional exploration of item-response level probabilities and their application will be saved for a future manuscript.

3.2 Sequential Probability Ratio Test (SPRT)

3.2.1 Overview

The sequential probability ratio test (SPRT) uses item-bank level probabilities and examinee responses to randomly drawn items from the item-bank to make classification decisions about an examinee. The following outlines the rule base and equations behind the SPRT approach to making classification decisions. Given a bank of items that assess mastery of a single learning objective, item-bank level probabilities can be either established empirically or set by the test administrator to form the SPRT rule base presented in Table 5. SPRT is based on the Classical Test Theory psychometric model and uses an item-bank where item-bank level probabilities have been established. In the absence of specific information about the probability that a given examinee belongs to a specific classification group equal prior probabilities are assumed.

Table 5. *SPRT Rule Base (adapted from Welch, 1997, p. 39)*

Rule	Description	Conditional Probability
1A	If the examinee is a *master* (M), the probability (P) of selecting an item that will be answered *correctly* (C) is .85	$P(C\|M) = .85$
1B	If the examinee is a *master* (M), the probability (P) of selecting an item that will be answered *incorrectly* (¬C) is 1 - .85 or .15	$P(\neg C\|M) = .15$
2A	If the examinee is a *nonmaster* (N), the probability (P) of selecting an item that will be answered *correctly* (C) is .40	$P(C\|N) = .40$
2B	If the examinee is a *nonmaster* (N), the probability (P) of selecting an item that will be answered *incorrectly* (¬C) is 1 - .40 or .60	$P(\neg C\|N) = .60$

During testing using the SPRT approach, an examinee is randomly administered an item from the item-bank, their response is evaluated, and the SPRT rule base is used to evaluate the probability that they are a master or a nonmaster. The test continues until either the probability ratio of the examinee being a master versus a nonmaster goes above or below specific threshold values or the maximum number of items has been administered to the examinee.

The probability that the examinee is a master is the product of the prior probability of mastery, the probability that a master would get the specific number of items correct, and the probability that a master would get the specific number of items incorrect (the numerator in equation 10). The probability that the examinee is a nonmaster is the product of the prior probability of nonmastery, the probability that a nonmaster would get the specific number of

items correct, and the probability that a master would get the specific number of items incorrect (the denominator in equation 10). The probability ratio that the examinee is a master versus a nonmaster can be calculated via equation 10.

The probability ratio is then compared to two threshold values (see equation 11) that are based on the acceptable error rates for making either a false mastery decision (α_{FM}) or a false nonmastery decision (β_{FN}) set by test administers. If the probability ratio is between the two thresholds then no-decision can be made and testing must continue. If the probability ratio is less than both thresholds then a nonmastery decision can be made and if it is greater than both thresholds then a mastery decision can be made.

$$PR = \frac{P_M P(C|M)^{\#C} P(\neg C|M)^{\#\neg C}}{P_N P(C|N)^{\#C} P(\neg C|N)^{\#\neg C}} \tag{10}$$

$$\frac{\beta_{FN}}{1-\alpha_{FM}} < PR < \frac{1-\beta_{FN}}{\alpha_{FM}} \tag{11}$$

Where:

PR = Probability ratio

$P(C|M)$ = Probability of a *correct* response given *mastery*

$P(\neg C|M)$ = Probability of an *incorrect* response given *mastery*

$P(C|N)$ = Probability of a *correct* response given *nonmastery*

$P(\neg C|N)$ = Probability of an *incorrect* response given *nonmastery*

P_M = Prior probability of *mastery*

P_N = Prior probability of *nonmastery*

$\#C$ = Number *correct* responses from examinee with unknown mastery status

$\#\neg C$ = Number *incorrect* responses from examinee with unknown mastery status

α_{FM} = Error rate established *a priori* for making false master decisions

β_{FN} = Error rate established *a priori* for making false nonmaster decisions

3.2.2 An Example

The following example of a single test administration uses the SPRT rule base presented in Table 5 and equal probabilities of a false mastery decision and false mastery decision ($\alpha_{FM} = \beta_{FN} = .025$) so that there is a 5% chance that SPRT will make an incorrect classification decision. Equation 11 is used to establish the threshold below which the probability ratio value would lead to a nonmastery decision ($PR < 0.026$) or above which the probability ratio would lead to a mastery decision ($PR > 39$).

Table 6. *Example of SPRT*

	R	Probability of R From:		Probability Examinee Is A:		*PR*	Test Decision
		Master	Nonmaster	Master	Nonmaster		
				.5	.5	1	
1	C	.85	.40	.680	.320	2.125	Continue
2	¬C	.15	.60	.347	.653	0.531	Continue
3	C	.85	.40	.530	.470	1.129	Continue
4	C	.85	.40	.706	.294	2.399	Continue
5	C	.85	.40	.836	.164	5.098	Continue
6	C	.85	.40	.915	.085	10.833	Continue
7	C	.85	.40	.958	.042	23.019	Continue
8	C	.85	.40	.980	.020	48.916	Master

Table 6 summarizes eight repetitions of the procedure of administering a random item, evaluating the examinee response, and determining if mastery or nonmastery decision can be made. Each repetition and row in Table 6 corresponds to the administration of a new item. For every item administered, Table 6 details the probability of the response (correct or incorrect) from a master and a nonmaster from the SPRT rule base established during calibration, the subsequent probability that the examinee is a master or nonmaster, the probability ratio (*PR*) from equation 10, and associated test decision based on comparing the *PR* to the upper threshold (> 39) and lower threshold (< .026) detailed in equation 11.

After administering eight questions the examinee has answered all but one item correctly and enough data has been collected to enable a mastery decision to be made so the test may terminate. Note that before any items have been administered the probability that

the examinee is a master is the same as the probability that the examinee is a nonmaster and that the probability ratio is 1.

After the first item is administered and the examinee responds correctly equation 10 is used to calculate the probability ratio with the number of correct responses equaling one and the number of incorrect responses equaling zero. The probably ratio after one item indicates that the examinee is over twice as likely to be a master versus a nonmaster but testing must continue since the probability ratio is less than the upper threshold. When the examinee responds incorrectly to the next question the probability ratio is updated and now the examinee is more likely to be a nonmaster than a master. Again, the probability ratio is still between the two threshold values so testing must continue. From this point on the examinee answers all the items administered correctly and, consequently, the probability ratio steady climbs until it passes the upper threshold and a classification decision of master can be made.

3.3 Expert Systems Enhanced SPRT with Random Item Selection (EXSPRT-R)

3.3.1 Overview

The EXSPRT-R is much like the SPRT in that (1) items are randomly selected from an item pool for administration and (2) the test terminates once the confidence in a classification decision reaches a specific threshold. The key difference is that EXSPRT-R applies expert systems thinking to apply item-level probabilities of a correct answer from specific classification groups in estimating the likelihood that an examinee belongs to a specific classification group. Table 7 presents the rule base for a specific fictional item number 63. The specific item parameter estimates for each item are applied in equation 12 to calculate the likelihood ratio of an examinee belonging to a particular classification group.

The test continues until so long as the likelihood ratio remains between the upper and lower thresholds defined in equation 13.

Table 7. *EXSPRT-R Rule Base for Item 63 (adapted from Welch, 1997, p. 47)*

Rule	*Description*	*Conditional Probability*
1A	If the examinee is a *master* (M) and item 63 is selected, the probability (P) of a *correct* response (C_{63}) is .89	$P(C_{63}\|M) =$ .89
1B	If the examinee is a *master* (M) and item 63 is selected, the probability (P) of an *incorrect* response ($\neg C_{63}$) is 1 - .89 or .11	$P(\neg C_{63}\|M) =$ .11
2A	If the examinee is a *nonmaster* (N) and item 63 is selected, the probability (P) of a *correct* response (C_{63}) is .65	$P(C_{63}\|N) =$ .65
2B	If the examinee is a *nonmaster* (N) and item 63 is selected, the probability (P) of an *incorrect* response ($\neg C_{63}$) is 1 - .65 or .35	$P(\neg C_{63}\|N) =$ .35

$$PR = \frac{P_M \prod_{i=1}^{n} P(C_i|M)^C P(\neg C_i|M)^{\neg C}}{P_N \prod_{i=1}^{n} P(C_i|N)^C P(\neg C_i|N)^{\neg C}} \quad (12)$$

$$\frac{\beta_{FN}}{1-\alpha_{FM}} < LR < \frac{1-\beta_{FN}}{\alpha_{FM}} \quad (13)$$

Where:

PR = Probability ratio

$P(C_i|M)$ = Probability of a *correct* response to item *i* given *mastery*

$P(\neg C_i|M)$ = Probability of an *incorrect* response to item *i* given *mastery*

$P(C_i|N)$ = Probability of a *correct* response to item *i* given *nonmastery*

$P(\neg C_i|N)$ = Probability of an *incorrect* response to item *i* given *nonmastery*

P_M = Prior probability of *mastery*

P_N = Prior probability of *nonmastery*

α_{FM} = Error rate established *a priori* for making false master decisions

β_{FN} = Error rate established *a priori* for making false nonmaster decisions

And:

$C = 1$, $\neg C = 0$ if item *i* is answered *correctly* by the examinee

Or:

$C = 0$, $\neg C = 1$ if item *i* is answered *incorrectly* by the examinee

3.3.2 An Example

The following example of a single test administration uses the item-level probabilities required for EXSPRT-R and uses equal probabilities of a false mastery decision and false mastery decision ($\alpha_{FM} = \beta_{FN} = .025$) so that there is a 5% chance that EXSPRT-R will make

an incorrect classification decision. Note that probability estimates are required for each item in the item bank for EXSPRT-R. Rather than list all probabilities associated with all possible items, Table 8 presents only the probabilities of specific responses to specific items. For example, the probability of a mastery responding incorrectly to item 63 is .11 and the probability of a nonmaster responding incorrectly to item 63 is .35. Equation 13 is used to establish the threshold below which the probability ratio value would lead to a nonmastery decision ($PR < 0.026$) or above which the probability ratio would lead to a mastery decision ($PR > 39$).

Table 8. *Example of EXSPRT-R*

	i	R	Probability of R To *i* From:		Probability Examinee Is A:		*PR*	Test Decision
			Master	Nonmaster	Master	Nonmaster		
					.5	.5	1	
1	63	¬C	.11	.35	.239	.761	.314	Continue
2	23	C	.81	.24	.515	.485	1.064	Continue
3	1	¬C	.08	.53	.138	.862	0.160	Continue
4	38	¬C	.02	.14	.024	.976	.025	Nonmaster

After administering four questions the examinee has answered all but one item incorrectly and enough data has been collected to enable a nonmastery decision to be made so the test may end. Note that before any items have been administered the probability that the examinee is a master is the same as the probability that the examinee is a nonmaster and that the probability ratio is 1.

After the first item is administered and the examinee responds incorrectly equation 12 is used to calculate the likelihood ratio. The likelihood ratio after one item indicates that the examinee is more likely to be a nonmaster versus a master but testing must continue since the likelihood ratio is greater than the lower threshold. When the examinee responds correctly to the next question the probability ratio is updated and now the examinee is slightly more likely to be a master than a nonmaster. Again, the probability ratio is still between the two threshold values so testing must continue. From this point on the examinee answers the two items administered incorrectly and, consequently, the probability ratio declines until it drops below the lower threshold and a classification decision of nonmaster can be made.

3.4 A Bayesian Statistics Perspective on Classical Test Theory Item Calibration

Bayesian statistics provide useful tools for understanding the tradeoff between reducing the uncertainty in probability estimates and increasing the item calibration sample size. Estimated probabilities for use in variable-length computer classification testing (VL-CCT) methods can be determined through the use of a beta distribution (Frick, 1989). The following introduces the basics of beta distributions drawing heavily on an introductory text by Schmitt (1969) and email correspondence with Dr. Theodore Frick. With few exceptions, terminology and abbreviations are from Schmitt. All abbreviations used are explained in appendix A.

3.4.1 The Probability Density Function of the Beta Distribution

The number of correct and incorrect responses from members of a specific classification group during the item calibration phase forms a unique *probability density function* of the beta distribution. For example, beta (* | 2, 3) is the probability density

function corresponding to two successes (correct responses) and three failures (incorrect responses) shown in Figure 1.

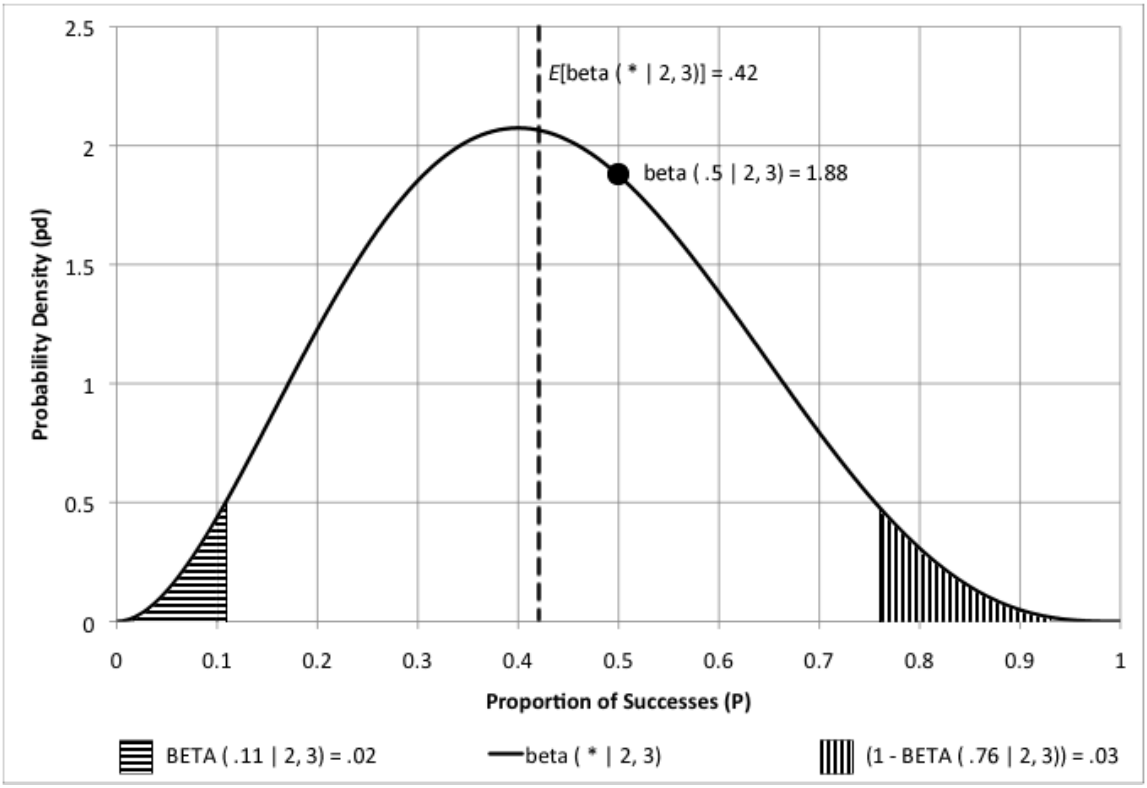

Figure 1. Probability Density Function for Two Successes and Three Failures

The horizontal axis represents the proportion of successes (*P*) with values ranging from zero to one. The vertical axis is the probability density (*pd*). The height of the probability density function represents the *pd* for a specific value of *P* and can be calculated using equation 14. For example, in beta (* | 2, 3) a *P* of .5 corresponds to a *pd* of 1.875 or in mathematical notation beta (.5 | 2, 3) = 1.88 (the large dot in Figure 1). Equation 1, introduced earlier, can be used to calculate the expected mean of a beta variable defined by *s* successes and *f* failures. The dashed line in Figure 1 represents the expected mean of the beta distribution in (.42).

$$beta(P|s,f) = \frac{(s+f+1)!}{s!\,f!} P^s (1-P)^f \tag{14}$$

The area under a portion of a probability density function is the probability (p) that the true proportion of successes (P) lies within a particular range of P. The area under a beta distribution from zero to any P value is called the *cumulative distribution function* and can be determined via numerical integration[1]. In Figure 1 the area under beta (* | 2, 3) from P=0 to P=.11 is filled with horizontal lines. Mathematical notation for the cumulative distribution function for P=.11 is BETA (0.0 to .11 | 2, 3) = .02. In other words, the p of the true P being between 0 and .11 in the probability density function defined by two successes and three failures is .02. The total area under the curve of a probability density function always equals one since zero and one bound all possible values of P and must correspond to a p of one. As such, the area under the probability density function curve to the right of any value of P is equal to one minus the value of the cumulative beta function at P. In Figure 1 the area under the function from P=.76 to P=1 filled with vertical lines equals a p of .03 (BETA (.76 to 1.0 | 2, 3) = .03).

3.4.2 Probability Density Function Variance, SD, and HDR

Several measures exist that may be applied to evaluate the preciseness of the expected mean proportion of successes (P), otherwise know as the beta mean (equation 1), associated with a given probability density function. The beta variance (equation 15) and beta standard

[1] Simpson's Rule is a proven numerical integration approach described by Schmitt (1969) for accurately estimating the area under any curve

deviation (SD) (equation 16) both provide measures of how spread out the associated probability density function is with *SD* being the more commonly reported. Both variance and standard deviation decrease as spread decreases and, as such, can be used as a measure of the precision of the estimated mean. It is important to note that both the variance and standard deviation equations are associated with the beta distribution not the normal distribution. While it is best practice to report *SD* values whenever means are provided, use of the beta mean in the VL-CCT literature has not included a pairing with the beta *SD* as a means of evaluating the uncertainty associated with estimated means.

$$Beta\ Variance = \frac{(s+1)(f+1)}{(N+2)^2(N+3)} \tag{15}$$

$$Beta\ Standard\ Deviation = \sqrt{\frac{(s+1)(f+1)}{(N+2)^2(N+3)}} \tag{16}$$

Where:

s = Number of successes

N = Number of observations $(s + f)$

f = Number of failures

An alternative measure of variance is associated with the *highest density region* (HDR), also known as a Bayesian confidence interval or a credible interval (Lee, 2004). HDR is an interval of *P* values associated with a specific fraction of the total probability (the area under the curve) such that every *P* value inside the interval has a higher *pd* than every *P* value outside the interval. The previous use of *P* values of .11 and .76 to illustrate areas under the probability density function was not by chance: they also represent the lower and upper bounds of the 95% HDR for beta (* | 2, 3). Recall, the area under the curve from a *P*

of 0 to .11 was .02 and .03 for area under the curve from a *P* of .76 to 1. Summing the two areas (.02 + .03) yields a *p* of .05, which is equal to the area outside the 95% HDR for beta (* | 2, 3). The 95% HDR tells us that for beta (* | 2, 3) the posterior probability that the true value of *P* is contained in the interval .11 to .76 is .95. The width of the HDR suggests how much confidence should be placed in an expected mean. In the case of beta (* | 2, 3), the width of the 95% HDR is .65 (.76 - .11 = .65) telling us that there is a large amount of uncertainty about the true value of *P*. The expected mean of *P* is only slightly more likely to be the true *P* than other values, some of which are substantially higher or lower.

Examining the *SD* leads to a similar conclusion regarding the confidence that should be placed in the mean. A *SD* of .175 is associated with beta (* | 2, 3) and the mean of .42 which means that we cannot have much certainty or confidence that the mean is a precise estimate. Fortunately, with additional observations the width of the 95% HDR can be narrowed, the *SD* decreased, and the uncertainty about true *P* reduced.

3.4.3 Sample Size and Kurtosis

More observations (i.e. successes and failures) yield probability density functions that are more peaked (lower variance), which serves to increase the *pd* over a narrower range of *P*. In other words, as sample size increases, the width of the associated HDR and *SD* decreases and the expected mean of *P* becomes more likely to be closer to the true value of *P*. Figures 2 through 7 illustrate how probability density functions become increasingly leptokurtic with more observations. Each presents two probability density functions with the black line representing data gathered from masters with a higher probability of responding correctly than nonmasters represented by the grey line.

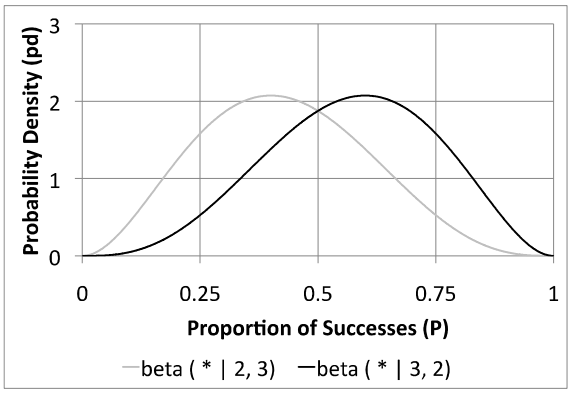

Figure 2. More Equal Ability Groups at 5 Observations

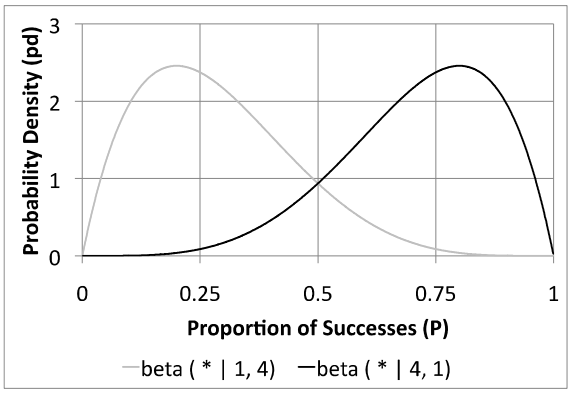

Figure 3. Less Equal Ability Groups at 5 Observations

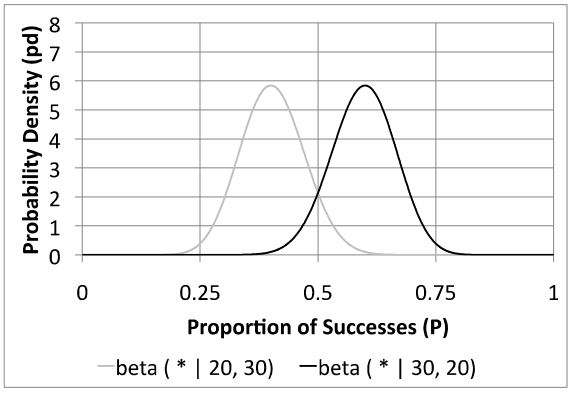

Figure 4. More Equal Ability Groups at 50 Observations

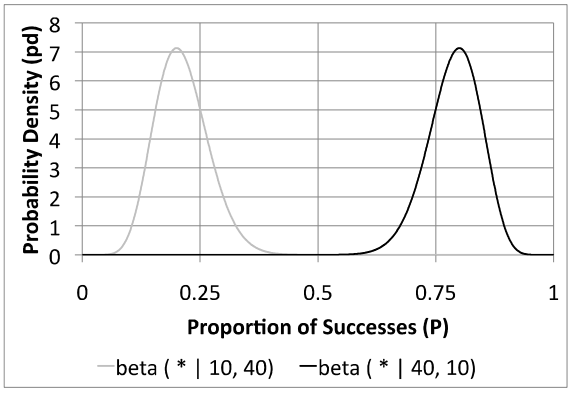

Figure 5. Less Equal Ability Groups at 50 Observations

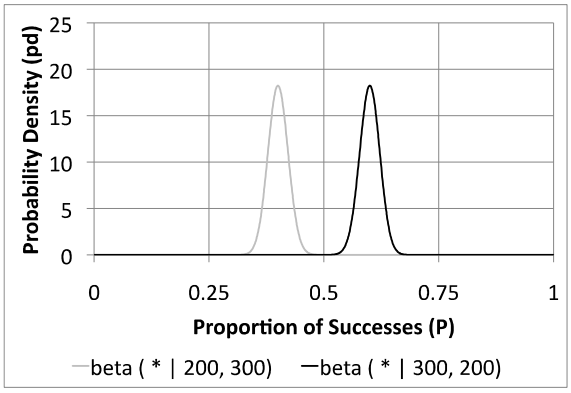

Figure 6. More Equal Ability Groups at 500 Observations

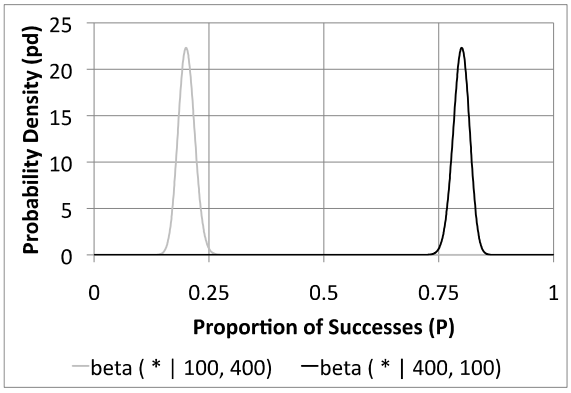

Figure 7. Less Equal Ability Groups at 500 Observations

Figures 2, 4, and 6 show probability density functions where masters respond correctly three times for every two incorrect responses and nonmasters respond correctly twice for every three incorrect responses. Similarly, Figures 3, 5, and 7 show probability density functions where masters respond correctly four times for every incorrect response and nonmasters respond correctly once for every four incorrect responses. While unlikely, consistent correct/incorrect response ratios are useful for showing how probability density functions differ when the sample size increases. Figures in the same row are based on the same sample size (5, 50, and 500 respectively). Note that the scale of the probability density axis changes to accommodate increasingly higher *pd* values associated with more observations.

Inspection of Figures 2 through 7 reveals three important points that have implications for item calibration. First, increasing the sample size results in more leptokurtic probability density functions with expected means of *P* that are more likely to be closer to the true *P*. The unsurprising implication is that larger calibration sample sizes will yield estimates of a probability of a correct response for a specific classification group that are more likely to be closer to the true value than estimates based on smaller calibration sample sizes. However, increasing the sample size comes at a cost: more examinees and more time both from examinees and administrators. Recall, Frick (1992) found that: (1) with 25 examinees per classification group sufficient classification accuracy could be obtained; (2) having more than 25 examinees per classification group did not substantially increase classification accuracy. It may be tempting to answer the question "How much information must be collected during item calibration to facilitate accurate classification decisions?" with

"25 examinees per classification group", however, a more nuanced answer is needed that accounts for differences in probability density functions associated with each classification group for reasons outlined next. Sample size is a coarse measure since it aggregates the number of correct and incorrect responses into a single number.

3.4.4 Beyond Sample Size

The second important point revealed by inspection of Figures 2 through 7 is that probability density functions based on the same sample size have different densities with those peaked nearer P=.5 being less dense than those peaked nearer P=0 or P=1. Probability density functions in Figures 3, 5, and 7 are all denser with higher peaks than their counterparts in Figures 2, 4, and 6 with the same sample size. The implication for item calibration is that after a fixed sample size items with a mean probability of a correct response near point five will be less likely to be close to the true probability than those items with a probability of a correct response closer to zero or one. In other words, item difficulty impacts accuracy of the estimated probability of a correct response and after a fixed sample size (e.g. 25) accuracy of estimates will vary according to item difficulty.

The third point to draw from Figures 2 through 7 is that the overlap between the nonmasters and masters probability density functions depends both on kurtosis and the difference in the proportion of successes. The area under both probability density functions decreases from Figure 2 to 4. The same tenfold increase in observations from Figure 3 to 5 eliminates the overlapping area entirely due, in large part, to the greater difference in abilities between nonmasters and masters. To understand the implications of this third point requires placing item calibration in the context of hypothesis testing which will be done next but, before moving on, the key takeaway from Figures 2 through 7 is that it cannot be assumed

that with a fixed sample size any item or set of items with any combination of probability density functions will be equally calibrated to enable accurate classification decisions. Instead, an approach is needed that takes unique probability density functions into account.

3.5 Item Calibration and Hypothesis Testing

The two probability density functions shown in each chart in Figures 2 through 7 are strikingly similar to representations used in statistical decision-making and suggest a framework for making decisions about when sufficient observations have been collected during item calibration that does account for the specific probability density functions. The following will first briefly review statistical decision-making and then present specific examples of how a statistical decision-making framework could be used to judge when enough observations have been made in the context of item calibration.

3.5.1 Type I and Type II Error

Statistical decision-making involves testing a null hypothesis that is either accepted or rejected based on the information available and is associated with two types of error. Type I error is incorrectly rejecting the null hypothesis when it is true. Type II error is incorrectly accepting the null hypothesis when it is false. The four possible results of hypothesis testing, including the two types of errors, are summarized in Table 9.

The probability of making a type I error is often designated alpha (α). Likewise, the probability of making a type II error is often designated beta (β). The symbol β will be used instead of the word "beta" to avoid confusion with the beta distribution. The value of α is typically not calculated, rather, statistical significance criterion is used to set α. Commonly used values for α are .05 and .01. The critical value, also know as the decision threshold (Schmitt, 1969), is used to judge if an observed measure supports or contradicts the null

hypothesis, is dependent on α because α must equal the probability of an observed measure being *larger* than the critical value while the null hypothesis is *true*. In turn, the value of β is dependent on the critical value since β is equal to the probability of an observed measure being *smaller* than the critical value while the null hypothesis is *false*. In other words, the value of α impacts the critical value which subsequently impacts the value of β. Consequently, decreasing α results in an increase in β. Sample size also influences β. Power analysis can be used to determine the minimum sample size that will enable an acceptable β for a given α and effect size (Cohen, 1988). Power is equal to 1 - β with .8 commonly used as an adequate level of power. How the value of α may impact decisions regarding if enough information has been collected during item calibration is addressed next.

Table 9. *Four Results of Hypothesis Testing*

	Decision	
Reality	Accept Null Hypothesis	Reject Null Hypothesis
Null Hypothesis is true	Correctly accept null hypothesis Probability = 1 - α	Incorrectly reject null hypothesis Probability = α Type I Error
Alternative Hypothesis is true	Incorrectly accept null hypothesis Probability = β Type II Error	Correctly reject null hypothesis Probability = 1 - β Power

3.5.2 Hypothesis Testing and Item Calibration Decisions

Two key decisions associated with item calibration can be informed through the application of the hypothesis testing approaches discussed above: (1) Does the item sufficiently discriminate between nonmasters and masters? (2) Are item parameter estimates sufficiently precise? The following will focus on the former. First, index of discrimination is introduced. Second, an example is provided to illustrate how calibration sample size can impact decision errors. Finally, a framework is presented based on hypothesis-testing for making decisions regarding if an item sufficiently discriminates between nonmasters and masters.

Index of discrimination, otherwise known as discrimination index, is an important measure of item quality. The index of discrimination for item *i*, represented by D_i, can be calculated by subtracting the probability of a nonmaster responding correctly to item *i* from the probability of a master responding correctly to item *i* (equation 15). Since probabilities range from zero to one the index of discrimination ranges from negative one to one. Table 10 presents guidelines on interpreting the index of discrimination put forward by Ebel (1972).

$$D_i = P(C_i|M) - P(C_i|N) \tag{17}$$

Where:

D_i = Index of discrimination for item *i*

$P(C_i|M)$ = Probability of a *correct* response to item *i* given *mastery*

$P(C_i|N)$ = Probability of a *correct* response to item *i* given *nonmastery*

Table 10. *Categorization of discrimination values adapted from Ebel (1972)*

Index of Discrimination	Item Evaluation
≥ .40	Very good items
.30 to .39	Reasonably good but possibly subject to improvement
.20 to .29	Marginal items, usually needing and being subject to improvement
≤ .19	Poor items, to be rejected or improved by revision

While it can be argued that the specific values of the index of discrimination associated with the very good, reasonably good, marginal, and poor items in Table 10 are somewhat arbitrary, two important points can be drawn from Ebel's categorization framework. First, high indices of discrimination (e.g. above .4) are associated with high quality items. The reason is that when masters are much more likely than nonmasters to respond correctly to an item then a correct or incorrect response to the item by an examinee provides more information that can be used in estimating an examinee's mastery status. Second, below a certain index of discrimination (e.g. below .19) items should either be rejected or revised. When masters are only slightly or not more likely than nonmasters to respond correctly to an item then a correct or incorrect response to the item by an examinee provides minimal information that can be used in estimating the examinee's mastery status. Further, it is standard practice to identify negatively discriminating items (those items that nonmasters are more likely to respond correctly to than masters) as problematic and remove them from the item-bank (Ebel, 1972). However, use of the index of discrimination alone can be misleading.

An item's index of discrimination is based on estimates of the probability of a correct answer from masters and nonmasters and the confidence in these probability estimates should influence the confidence in the associated index of discrimination. Consider Figures 2, 4, and 6 that show mastery and nonmastery beta distributions for an item at five, 50, and five 500 observations. Each of these is based on the same ratio of correct and incorrect responses from masters and nonmasters and their associated indices of discrimination (calculated using equations 6, 8, and 17) are all less than .2 thus qualifying them as *poor* items according to Table 10. On the other hand, Figures 3, 5, and 7 are all associated with indices of discrimination greater than .4 thus qualifying them as *very good* items in Table 10.

However, the index of discrimination hides the number of observations it is based on. Figures 6 and 7 are based on one hundred times more observations than Figures 2 and 3 and, as such, more confidence can be placed in the associated estimates of the probabilities of nonmasters and masters responding correctly to the item and the resulting index of discrimination. Making decisions regarding if an item sufficiently discriminates between nonmasters and masters with too few observations contributes to making two unique types of errors: (1) falsely identifying an item as *problematic* when in reality it is not and (2) falsely identifying an item as *not problematic* when in reality it is.

The two types of errors just identified are examples of type I and type II errors presented in Table 9. If the null hypothesis is "the probability of a master responding correctly is *less than or equal to* the probability of a nonmaster responding correctly" then existing hypothesis testing approaches can be applied that test the null hypothesis under conditions where acceptable levels of error are defined *a priori*. The following presents two

examples of testing this null hypothesis with the same data but with different values for α (the probability of making a type I error).

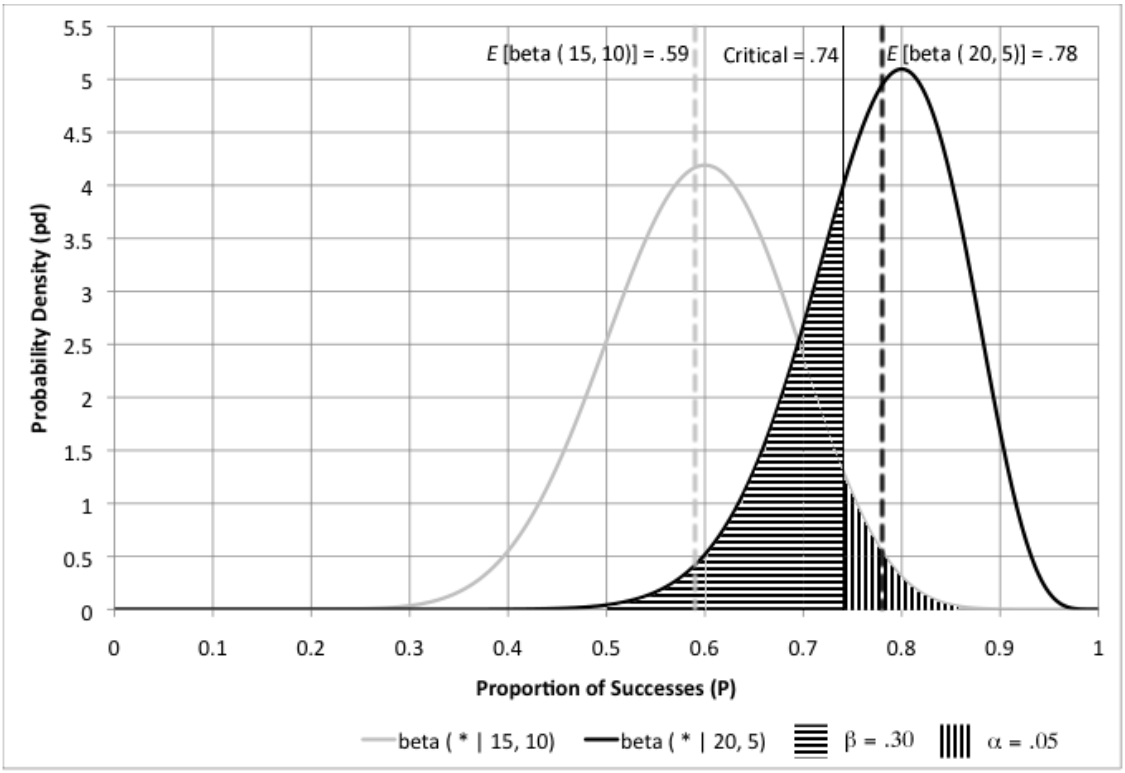

Figure 8. Difference between beta (* | 15, 10) and beta (* | 20, 5) at alpha = .05

Figure 8 presents nonmaster and master, probability density functions, beta (*, 15, 10) and beta (*, 20, 5), along with the critical value and β that would result from setting α to .05. When α is set to .05 then the critical value is the point where five percent of the area under the nonmaster probability density function lies to the right. Numerical integration using Simpson's Rule determines the critical point to be at *P*=.74. Shading with vertical bars denotes the α area of .05 in Figure 8. Simpson's rule establishes the area under the masters curve to the left of *P*=.74 to be β=.30 (represented in Figure 8 via shading with horizontal bars). If we use the expected mean of *P* for the masters probability density function as the observed value then via equation 2 the observed value equals .78. The observed value (.78) is

larger than the critical value (.74) so we could reject the null hypothesis and conclude that the probability of a master responding correctly is significantly larger (at the $\alpha = .05$ level) than the estimated probability of a nonmaster responding correctly (.59). However, the power of this test ($1 - \beta$) is .70, below the value commonly identified as adequate (.8).

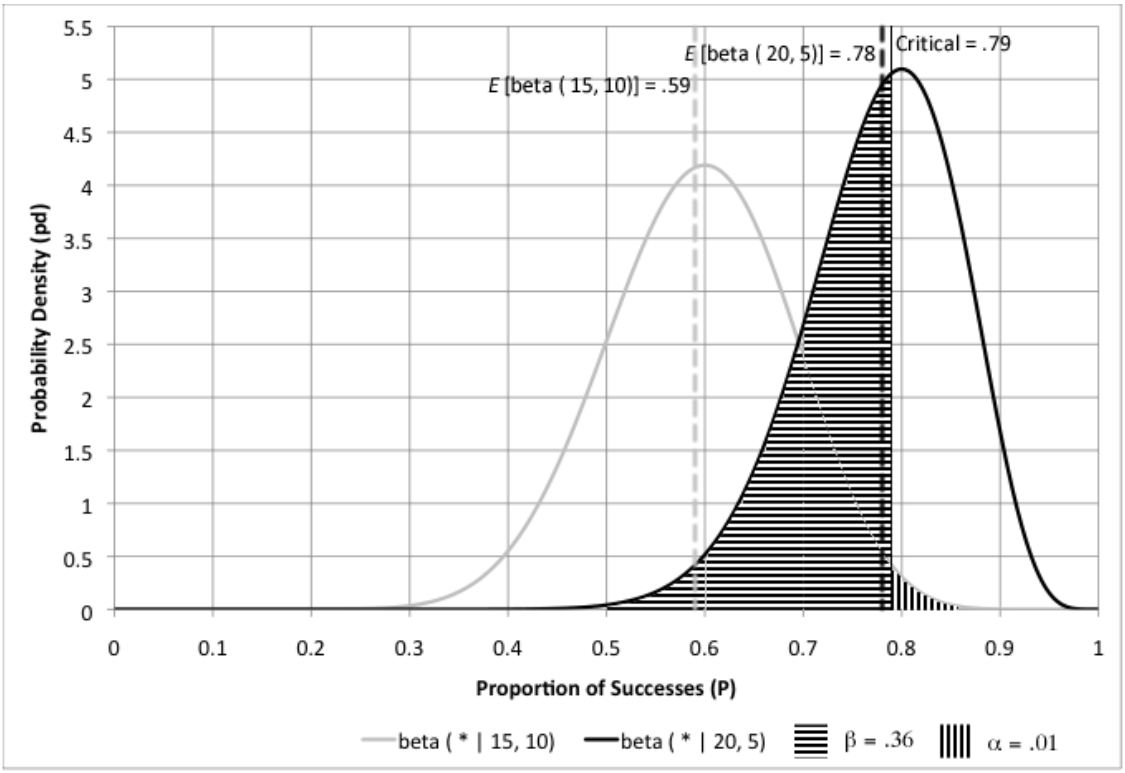

Figure 9. Difference between (* | 15, 10) and beta (* | 20, 5) at alpha = .01

Figure 9 presents the same probability density functions for nonmasters and masters but illustrates how an α of .01 would result in the null hypothesis being accepted rather than rejected. When α is .01 the critical value of *P* increases to .79 and β increases to .36. Consequently, the expected mean of *P* for the masters probability density function (.78) is smaller that the critical value (.79) so the null hypothesis could be accepted: no significant differences (at the $\alpha = .01$ level) in the estimated probability of a correct response from a master and a nonmaster. Setting α to .01 results in a different conclusion from that arrived at

by setting α to .05. Furthermore, decreasing α from .05 to .01 results in an increase in β from .30 to .36 and an increase in the risk of committing a type II error (lower power).

Clearly, decisions regarding the thresholds of α and β have implications for reaching conclusions regarding when items sufficiently discriminate between nonmasters and masters. However, unlike other types of statistical decision-making, guidelines regarding what thresholds of α and β enable efficient and accurate decisions regarding if an item sufficiently discriminates between nonmasters and masters do not exist. Likewise, statistical methods for determining the degree that one beta distribution differs from another have not been established. Consequently, researchers and test administrators are left to rely on less precise guidelines based on sample size until practical recommendations are established.

Figures 8 and 9 are based on testing the null hypothesis "the probability of a master responding correctly is *less than or equal to* the probability of a nonmaster responding correctly". Evidence that the probability of a master responding correctly is sufficiently greater than the probability of a nonmaster responding correctly supports (1) rejecting the null hypothesis and (2) the inclusion of the item in the item-bank. However, efficiently identifying poor items is also important. The same approach described above may be used to test a different but related null hypothesis, namely, "the probability of a master responding correctly *is greater than or equal* to the probability of a nonmaster responding correctly". Evidence that the probability of a master responding correctly is sufficiently less than the probability of a nonmaster responding correctly supports (1) rejecting the null hypothesis and (2) the exclusion of the item from the item-bank. Efficient identification and removal or pruning of problematic items from the item-bank can prevent additional time and resources being wasted in a vain attempt to calibrate an item that is highly likely to be problematic.

Before moving on to the proposed Automatic Racing Calibration Heuristics approach, the method for answering the second key question associated with item calibration must be addressed. When are item parameter estimates sufficiently precise? The previous discussion on highest density regions (HDR) and standard deviations of beta distributions suggests an answer: continue calibrating items until the variance of both the nonmaster and master beta distributions decreases to a point where it reaches a pre-established threshold that has been associated sufficient preciseness. However, much like the previous discussion about criteria for determining when one beta distribution mean is sufficiently different from another, criteria for determining when beta distribution means are sufficiently precise have not been established that lead subsequent testing to make classification decisions within pre-established error rates.

3.6 Automatic Racing Calibration Heuristics (ARCH)

3.6.1 Overview

Automatic Racing Calibration Heuristics (ARCH) applies statistical hypothesis testing techniques to efficiently address item calibration questions during online testing. ARCH pits SPRT and a slightly modified version of EXSPRT-R, labeled M-EXSPRT-R, against each other in a *race* to make accurate classification decisions about examinees using the fewest number of items. The simultaneous application of multiple VL-CCT approaches to was inspired by methods applied by Welch & Frick (1993) to compare SPRT and EXSPRT-R approaches during live testing. Initially, only item-bank level parameter estimates must be established for SPRT. Item level parameter estimates used by M-EXSPRT-R are not available at the outset. After each classification decision, ARCH: (1) *automatically* uses the additional response data gathered during the *online* test to update associated item *calibration*

parameter estimates and (2) applies a set of *heuristics* to determine if any items are sufficiently calibrated for use with M-EXSPRT-R. Through the course of online testing more items become sufficiently calibrated for use with M-EXSPRT-R, thus increasing the chances that M-EXSPRT-R will be able to make classification decisions more efficiently than SPRT. In short, tests become more efficient as testing progresses.

3.6.2 Measured-EXSPRT-R (M-EXSPRT-R)

M-EXSPRT-R is a VL-CCT approach that operates in nearly the same way as the EXSPRT-R approach presented earlier. The key difference is that M-EXSPRT-R only uses item parameter estimates for items that are sufficiently calibrated. An item is sufficiently calibrated if it meets specific discrimination and precision criteria. The M in M-EXSPRT-R represents the *measured* use of item parameter estimates. During the initial stages of a test using the ARCH approach few if any items in the item-bank will be sufficiently calibrated. As such, M-EXSPRT-R will have little or no data on which to make decisions about mastery. However, SPRT, the competing VL-CCT approach, relies on item-bank level parameter estimates that are available and can make decisions about examinees. The SPRT decisions generate additional item calibration data that lead to more items that are sufficiently calibrated for use with M-EXSPRT-R.

3.6.3 Design Components

ARCH requires a reasonably large item-bank calibrated at the item-bank level in order to start. The item-bank must be large enough to ensure desired item exposure rates are met. Only the four item-bank level parameter estimates used by SPRT (equations 6, 7, 8, and 9) must be established before ARCH can begin. The item-bank level parameter estimates

used by SPRT can either be set by test administrators or based on empirical data from a calibration phase involving a fixed number of nonmasters and masters.

3.6.4 Heuristics

ARCH applies a set of calibration heuristics every time a classification decision is made about an examinee by either SPRT or M-EXSPRT-R to evaluate any of the items that were used in making the classification decision can be accepted for use with M-EXSPRT-R. The heuristics are summarized in Table 11. For each item four yes or no questions are asked: (1) Is the number of times item *i* has been administered during calibration (n_i) greater than or equal to the calibration administration maximum (n_{max})? (2) For item *i*, is the probability of a correct response given *mastery* ($P(C_i|M)$) sufficiently *less than* the probability of a correct response given *nonmastery* ($P(C_i|N)$)? (3) For item *i*, is the probability of a correct response given *mastery* ($P(C_i|M)$) sufficiently *greater than* the probability of a correct response given nonmastery ($P(C_i|N)$)? (4) Are the estimates of the probability of a correct response from a nonmaster and the probability of a correct response from a master sufficiently precise? Note that while all but the first question can be stated as a hypothesis testing or statistical estimation problem, the word "sufficiently" has been intentionally used in the heuristic questions instead of "significantly" to keep the focus on establishing practical heuristic criteria that lead to accurate and efficient performance of M-EXSPRT-R.

Table 11. *Item-Level Calibration Heuristics Decision Table for ARCH*

<table>
<tr><td rowspan="5">$n_i \geq n_{max}$?</td><td>Yes</td><td colspan="5"></td><td>Reject i. Stop calibration on i.</td></tr>
<tr><td rowspan="4">No</td><td rowspan="4">Is $P(C_i|M)$ Sufficiently Less Than $P(C_i|N)$?</td><td>Yes</td><td colspan="3"></td><td>Reject i. Stop calibration on i.</td></tr>
<tr><td rowspan="3">No</td><td rowspan="3">$P(C_i|M)$ Sufficiently Greater Than $P(C_i|N)$?</td><td rowspan="2">Yes</td><td rowspan="2">$P(C_i|M)$ & $P(C_i|N)$ Sufficiently Precise?</td><td>Yes</td><td>Accept i. Stop calibration on i.</td></tr>
<tr><td>No</td><td rowspan="2">No Decision. Continue calibration on i.</td></tr>
<tr><td>No</td><td colspan="2"></td></tr>
</table>

Where:

n_i = Number of times item *i* has been administered during calibration

n_{max} = Maximum administrations for any item during calibration

$P(C_i|M)$ = Probability of a *correct* response to item *i* given *mastery*

$P(C_i|N)$ = Probability of a *correct* response to item *i* given *nonmastery*

The answers to each of these questions are used to select one of three possible outcomes for the given item: (1) Reject the item for use with M-EXSPRT-R and stop further calibration on the item; (2) Accept the item for use with M-EXSPRT-R and stop further calibration on the item; (3) Make no decision and continue calibration on the item. Initially no items will have been accepted for use with M-EXSPRT-R. However, as the ARCH approach continues calibrating, all items in the item-bank will be rejected or accepted for use

with M-EXSPRT-R. The answers to the yes or no questions and the outcome selected depend on: (1) what criteria are used to judge if the differences between $P(C_i|M)$ and $P(C_i|N)$ are sufficient and (2) what criteria are used to judge if estimates of $P(C_i|M)$ and $P(C_i|N)$ are sufficiently precise. The first of two studies of this dissertation, Monte Carlo ARCH operationalization and evaluation with historical COM test data, established specific ARCH heuristic criteria through identifying appropriate statistics and determining associated threshold values that enabled the M-EXSPRT-R component of the ARCH approach to make efficient classification decisions within *a priori* error rates.

3.6.5 An Example

Let's assume we start with an item-bank calibrated for use with SPRT and that the item-bank level parameter estimates are the same as those used in the example used to explain SPRT [$P(C|M) = .85$; $P(C|N) = .40$; $P(\neg C|M) = .15$; $P(\neg C|N) = .60$]. Suppose further, the ARCH approach has sufficiently calibrated a few items for use with M-EXSPRT-R. As with the EXSPRT-R example the specific probabilities of specific responses to specific items are provided in Table 13 rather than listing all four probabilities for all possible items. Like the previous examples of SPRT and EXSPRT-R, Table 13 illustrates the administration of a test to a single examinee. However, Table 13 demonstrates how both SPRT and M-EXSPRT-R approaches operate in parallel in a race to make a classification regarding the examinee using the data available to each approach.

After administering seven randomly selected items to the examinee only M-EXSPRT-R is able to make a classification decision about the examinee despite not being able to use three of the seven responses in it's decision making. Both SPRT and M-EXSPRT-R are able to use the responses to the first two items administered (items 63 and 23) to calculate

corresponding probability and likelihood ratios. When item 28 is administered only SPRT can use the examinee response to update the probability ratio. M-EXSPRT-R, on the other hand, does not yet know enough about item 28 to use the associated item-level parameter estimates to update the likelihood ratio that consequently remains at 1.064.

Item 28 represents an item that has not yet met the criteria defined by the calibration heuristics in Table 11 – item 28 has been neither accepted nor rejected for use by M-EXSPRT-R and is still in the process of being calibrated. Item 87 and 11 are also not sufficiently calibrated for use with M-EXSPRT-R so the examinee responses to these items are not incorporated into the M-EXSPRT-R likelihood ratio. However, since SPRT uses item-bank level parameter estimates, the examinee responses to all items are used to update the probability ratios.

Table 12. *Example of Racing SPRT and M-EXSPRT-R*

	i	R	Probability of R From:		*SPRT PR*	SPRT Test Decision	Probability of R To *i* From:		*M-EXSPRT-R PR*	M-EXSPRT-R Test Decision
			Master	Nonmaster			Master	Nonmaster		
					1				1	
1	63	¬C	.15	.60	0.250	Continue	.11	.35	.314	Continue
2	23	C	.85	.40	0.531	Continue	.81	.24	1.064	Continue
3	28*	¬C	.15	.60	0.133	Continue	-	-	1.064	Continue
4	1	¬C	.15	.60	0.033	Continue	.08	.53	0.160	Continue
5	87*	C	.85	.40	0.071	Continue	-	-	0.160	Continue
6	11*	C	.85	.40	0.150	Continue	-	-	0.160	Continue
7	38	¬C	.15	.60	0.037	Continue	.02	.14	.025	Stop: Nonmaster

* Item level calibration is not complete so M-EXSPRT-R cannot use associated item parameter estimates

CHAPTER IV. METHODS

This dissertation is made up of two related studies. Both examine RQ2 and RQ3 using the same analytical methods, however, they do so using different participants and data collection methods. Both studies evaluate the ARCH approach but only the first study had the additional purpose of operationalizing the ARCH concepts presented in the previous chapter and establishing the ARCH calibration criteria and thresholds that provide answers to RQ1. A one-page summary of the two studies, participants, research questions, and methods applied is presented in Table 13.

The first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data, provided preliminary answers to each of the three research questions: (RQ1) When is an item sufficiently calibrated? (RQ2) How accurate is ARCH? (RQ3) How efficient is ARCH? The second study, ARCH evaluation with new IU plagiarism test examinees, used the item calibration criteria established from the first study (RQ1) to determine if the preliminary answers RQ2 and RQ3 from the first study held true in the live online testing context associated with a new adaptive version of the Indiana University (IU) plagiarism test.

The organization of this chapter is consistent with Table 13. The following elaborates on Table 13, row by row, from the top to the bottom. The two main sections of this chapter correspond to the two studies and the associated rows of Table 13. The structure of each section is the same – participants associated with the phase are described, followed by a discussion of the methods of data collection and analysis in the context of specific research questions.

Table 13. *Methods Summary*

Study	Participants	Research Questions	Data Collection Method	Analytic Method
I: Monte Carlo ARCH operationalization and evaluation with historical COM test data	Examinees (*N*=104) from a previous study (Frick, 1992) who responded to an 85-item test. Participants came from 2 sections of a graduate course, one undergraduate course, and a few volunteer recruits from IU's main library	1. When is an item sufficiently calibrated?	*Monte Carlo simulations with historical COM test data:* Historical Examinee item responses were used to simulate test administrations using a variety of ARCH criteria	*Repeated simulations* to identify specific ARCH criteria that are likely to achieve sufficient test classification accuracy while hastening deployment of M-EXSPRT-R to improve efficiency
		2. How *accurate* is ARCH?	*Monte Carlo simulations with historical COM test data:* Recording of resulting classification decisions and test lengths from simulated test administrations using established ARCH criteria.	*Tests* to determine if ARCH classification decisions deviate significantly from total test classification decisions
		3. How *efficient* is ARCH?		*Friedman Test and Post Hoc Testing* to determine if mean test lengths of ARCH, SPRT, and EXSPRT-R tests are significantly different
II: ARCH evaluation with new IU plagiarism test examinees	Volunteers (*N*=5,729) from the thousands who take the IU plagiarism test	2. How *accurate* is ARCH?	*Live Testing:* Examinee item responses were collected. For each examinee, the number of items required to make a classification decision and the decision itself were recorded for each testing method.	*Chi-squared Tests* to determine if ARCH, calibrated using the ARCH criteria established in the first study makes classification decisions that deviate significantly from decisions made by EXSPRT-R calibrated with 50 masters and 50 nonmasters
		3. How *efficient* is ARCH?		*Friedman Test and Post Hoc Testing Repeated Measures One-Way ANOVA and Post Testing* to determine if mean test lengths of ARCH, SPRT, and EXSPRT-R tests are significantly different

4.1 Monte Carlo ARCH Operationalization and Evaluation with Historical COM Test Data

The first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data, involves test re-enactments via Monte Carlo computer simulations using historical test data from a previous study (Frick, 1992) in order to both operationalize and evaluate ARCH. In addition to providing preliminary answers to each of the three research questions, the ARCH criteria necessary for accurate and efficient testing established in the first study were used for ARCH in the second study.

Simulation Method

Monte Carlo simulations were conducted to reenact tests. A combination of JavaScript and HTML was used to create the various types of simulations that are described in greater detail in the results chapter. Common to all Monte Carlo simulations is the method for how items were selected for administration and the examinee response determined. In all cases, random item selection without replacement was used. In other words, items were randomly selected for a given test for a particular examinee such that no item was administered more than once on a particular test. The examinee response to an item was always drawn from their actual response to the item in the historical data. This means that if examinee *A* responded to item 32 incorrectly in the historical data then an incorrect response would be recorded anytime a simulation involving examinee *A* involved administration of item 32.

The basis for the methods used in this study is the ability to use historical data of examinee responses to specific questions to simulate various testing approaches and record results including the test length, the classification decision, and if the classification decision agrees with the classification decision made using all the test items. Each of the testing approaches being examined uses random item selection, which is advantageous in terms of the large volume of different simulations possible. The same examinee can be tested multiple times using the same testing approach with random item selection since there are over 2.8×10^{128} or 85! different ways to select 85 test items. Consequently, multiple samples can be generated from the historical data for the purposes of addressing each of the research questions.

Historical COM Test Data

The historical test datum on which the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data, was based include responses from 104 participants on a computer literacy test, subsequently referred to as the COM test, comprised of 85 items. According to Frick (1992), the examinees came from three sources: (1) two sections of a graduate course on the use of computers in education accounted for nearly half of the examinees; (2) volunteers from an undergraduate course on how to use computers for non-education majors constituted the second largest source of examinees; and (3) only a few examinees were volunteers recruited at Indiana University's main library.

The test on the subject of how computers work was comprised of 85 items that included "about half multiple-choice, one-fourth binary choice, and one-fourth fill-in type

questions" (Frick, 1992, p. 203). Analysis of the responses enabled a cut score between nonmasters and masters of 72.5% to be established and resulted in dividing the examinees into 28 nonmasters (27%) who responded correctly to less than 72.5% of the items and 76 masters (73%) who responded correctly to 72.5% or more of the items. The Cronbach alpha for the test was .94 and the mean total correct for all examinees was 79%.

Familywise Error Rate

The familywise error rate for this study was set to the often used value of $\alpha = .05$. As there were five hypothesis tests being performed, the *p*-value used for testing each of the hypotheses was set to $p = .01$. RQ2 included performing statistical analysis on three hypotheses tests: (1) Do overall error rates differ significantly across the four testing algorithms? (2) Do rates of false nonmastery differ significantly across the four testing algorithms? and (3) Do rates of false mastery differ significantly across the four testing algorithms? The remaining two hypothesis tests were associated with RQ3: (3) Do mean test lengths differ significantly across the four testing algorithms? and (4) Do no-decision rates differ significantly across the four testing algorithms?

Post-hoc testing associated with any of the four hypothesis tests listed above to look for pairwise differences between the four testing algorithms requires that the $p = .01$ value be divided further. Four testing algorithms equate to the following six pairwise comparisons: (1) SPRT and ARCH pre-calibration; (2) SPRT and ARCH post-calibration; (3) SPRT and EXSPRT; (4) ARCH pre-calibration and ARCH post-calibration; (5) ARCH pre-calibration and EXSPRT; (6) ARCH post-calibration and

EXSPRT. Applying the Bonferroni correction results in dividing $p = .01$ by six, which yields the small value of $p = .0017$ that was used in post-hoc analysis.

Since ARCH pre-calibration uses the SPRT testing algorithm, it could be argued that comparing ARCH pre-calibration and SPRT is not warranted. However, test reenactments in this study enabled ARCH pre-calibration and SPRT to independently select items randomly for administration. In other words, an ARCH pre-calibration reenactment and a SPRT reenactment for the same examinee would be expected to differ slightly in both accuracy and efficiency due to the random selection of different items. These differences were not expected to be significant since they were due to random chance, so comparisons between ARCH pre-calibration and SPRT provided a way of confirming that ARCH pre-calibration performed as expected (i.e. does not differ significantly from SPRT).

Goodness-of-fit tests conducted in study 1 were not included in the overall familywise error rate and individual hypothesis testing *p* values. Unlike the other statistical tests of significance conducted in study 1, lack of a significant difference between the error rates observed with the ARCH algorithm and the expected *a priori* error rates represents a discovery. Consequently, keeping the *p* value at .05 is a conservative approach, as the probability of finding a significant result is greater than if *p* was decreased.

4.1.1 ARCH Calibration Sufficiency (RQ1) Data Collection & Analytic Methods

RQ1 sought to identify specific criteria that enable ARCH to reliably predict when an item has been sufficiently calibrated to enable testing to make classification decisions within *a priori* error rates. The data collection method used to answer RQ1

involved using the historical test data in the first stage of the computer simulations planned for the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data.

ARCH item calibration and testing were simulated with a range of ARCH criteria to determine what specific combinations of measures and associated thresholds predictably led to sufficiently accurate testing while hastening the deployment of the more efficient M-EXSPRT-R testing algorithm via ARCH. Classification decisions made for each examinee for each simulation were compared to the true mastery classification status of an examinee to determine the frequency of correct and incorrect classification decisions. Comparison of the examinee total score on the COM test to the mastery cutoff indicated their mastery status.

For example, recall that Frick's (1992) study based on COM test data found that increasing the calibration sample from 25 examinees per classification group to 50 examinees per classification group did not lead to substantial gains in either efficiency or classification accuracy. This finding informed the simulations in the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data, aimed at answering RQ1. Using the upper bound of 50 examinees per classification group enabled the upper bounds (most conservative or strict bounds) of various calibration criteria to be established.

Once the baselines were established, the values of ARCH criteria were systematically adjusted in subsequent simulations of calibration and testing to establish what impact various criteria thresholds would have on the probability of obtaining *a priori* error rates. Extremely conservative ARCH criteria thresholds meant that resulting

testing based on ARCH calibration would meet *a priori* classification error rates; however, these ARCH conservative criteria also served to delay the deployment of the more efficient M-EXSPRT-R. The goal was to determine the ARCH heuristic criteria and associated thresholds that represent a good, not necessarily optimal, tradeoff between ensuring classification accuracy and the timely deployment of M-EXSPRT-R.

Each set of possible ARCH calibration criteria was evaluated using 2,080 simulated ARCH tests, which equates to giving each of the 104 examinees an ARCH test twenty times. Over the course of the 2,080 tests, ARCH switched from SPRT-based testing to the racing approach where SPRT and EXSPRT were used in parallel, with the first one able to make a classification decision winning the race. The timing of the switch from SPRT to the racing approach depended on the ARCH calibration criteria and the random selection of both examinees and items. Both items and examinees were selected without replacement to ensure even use of both during the simulations.

4.1.2 ARCH Accuracy (RQ2) Data Collection & Analytic Methods

The second research question addresses the accuracy of testing classification decisions that result from application of the ARCH approach. Again, the method for collecting the required data involved Monte Carlo computer simulations using historical COM test data. Historical COM test data were used to perform: (1) multiple rounds of ARCH calibration and testing using the ARCH criteria established during the earlier Monte Carlo computer simulations; (2) SPRT testing; and (3) EXSPRT calibration and testing.

Data Generation Monte Carlo Simulations

To ensure independence of observations necessary for subsequent statistical analysis for both RQ2 and RQ3, the unit of analysis was the examinee. After conducting numerous simulations, each examinee was associated with test accuracy (RQ2) and test efficiency (RQ3) measures for four testing algorithms: SPRT, ARCH pre-calibration, ARCH post-calibration, and EXSPRT. All algorithms were permitted to continue test simulations until either a classification decision could be made or all the items in the item-pool had been deployed. The classification decisions made using each of the four testing algorithms were compared to classification decisions made using the total-test decision that served as the examinee's true mastery state.

Instead of limiting the data being analyzed to a single simulated test of each of the four testing algorithms, 50 simulated tests were conducted for each of the four testing algorithms, resulting in mean test accuracy and test efficiency statistics for each of the 104 examinees. A single test can be prone to random variations, whereas 50 tests are more likely to provide stable results. For example, a single simulated SPRT test for a specific examinee ended after only four items but 50 simulations of the SPRT test with the same examinee reveals that, on average, the SPRT test length was over 31 items.

Testing Algorithm Settings & Calibration

Calibration requirements differed across each of the four algorithms used in each of the 50 test simulations per examinee, however, *a priori* error rates were all set to false nonmaster rate = false master rate = 2.5% for a total error rate of 5%, and the prior probability of mastery was set to .5 for all algorithms. SPRT does not rely on the collection of data to calibrate test items but instead relied on overall probability of a

correct answer from a master (.85) and a nonmaster (.60) to make classification decisions with specific values drawn from earlier studies of COM test data (Frick, 1989). EXSPRT was calibrated with all available test data from all 104 examinees (28 nonmasters and 76 masters), where those who had a total score of greater than or equal to 72.5% were considered masters. ARCH calibration requirements are more complex.

ARCH depends on calibration during testing and evaluation against specific calibration criteria established via answering RQ1. For RQ2 and RQ3 multiple rounds of ARCH simulations were conducted with sets of simulated tests being randomly drawn for each of the 104 examinees both before and after ARCH calibration criteria had been met. A total of seven rounds of ARCH simulations were conducted before 50 ARCH pre-calibration and 50 ARCH post-calibration sets were available. ARCH reached the calibration criteria established as part of answering RQ1 at different points in each ARCH simulation. For example, ARCH shifted to the racing testing approach at tests 726, 768, 753, 728, 710, 733, and 782 in the seven ARCH simulations used to generate data for this study. Post-calibration ARCH tests were randomly drawn from a set of tests after calibration criteria had been met that reflected the number of tests it took to reach calibration criteria.

Analytic Methods

Pearson's chi-squared test, otherwise known as the chi-squared goodness-of-fit test, can be used to determine if an observed distribution of a categorical variable (e.g. test decision) into specific groups (e.g. correct or incorrect decision) follows an expected distribution established *a priori*. Two of the assumptions associated with Pearson's chi-

squared test have consequences on the type of analysis that can be done with the COM test data. First, the Pearson's chi-squared test is based on the assumption of independence of observations, which means that it would be inappropriate for two observations (i.e., test simulation decisions) used in the same Pearson's chi-squared test to be based on the same examinee, as this would violate independence of observations.

A second assumption of the Pearson's chi-squared test is that there must be a minimum of five expected frequencies in each categorical variable group. The *a priori* false nonmaster and false master error rates are both 2.5% so with 104 examinees the expected frequency of each would be 2.6, which is less than the five necessary to meet the assumption. Consequently, a nonasymptotic method was applied during chi-squared testing that involves 10,000 Monte Carlo simulations to compute an exact *p* value with a confidence of 99% that is reliable in situations where there are many cells with counts less than five (Mehta & Patel, 1989).

All 100 chi-squared tests were conducted with $p \leq .05$, indicating a significant lack of fit between observed error rate in the given set of 104 tests and the expected *a priori* error rate. The chances of making a type I error – incorrectly finding a significant result – goes up with the number of hypothesis tests conducted. With 100 chi-squared tests and $p \leq .05$ indicating a significant lack of fit, the chances are very high ($1 - 0.95^{100} = 0.99$) of finding at least one significant lack of fit when, in fact, the fit is not significantly different. However, since the desired result is to not find a significant lack of fit, keeping the *p* value at .05 instead of lowering it represents a conservative approach.

Power analysis, using the software application G*Power (Faul et al., 2007) indicated that the 104 examinees in the historical COM test data were sufficient to conduct the goodness of fit tests able to detect large effect sizes. Power analysis using an effect size of .37, a chance of making a type I error of .025, a power of .8, and two degrees of freedom indicates that a sample size of only 93 is required for the goodness of fit tests, meaning that the 104 examinees in the existing sample were adequate. An effect size of .37 is considered a large effect based on guidelines for the social sciences provided by Cohen (1988). The use of two degrees of freedom reflects that classification decisions were constrained to three options: mastery, nonmastery, and no decision.

Friedman Tests and Wilcoxon Signed-Rank Post-Hoc Testing

The nonparametric Friedman Test was used to see if overall, false nonmaster, and false master rates differed significantly according to the testing algorithm used with post hoc analyses via Wilcoxon signed-rank tests used to determine which specific pairwise comparisons were significant.

Proportion Reduction in Error

Analysis included calculation of the proportion reduction in error. The Proportion Reduction in Error (see Rudner, 2009, p. 7) is a useful method for comparing accuracy of classification decisions since it takes into account the probability that a decision could be accurate simply by chance, which goes down as the number of possible classifications increases. Accounting for percent accurate by chance is important to allow for comparisons between approaches that have different classification numbers; otherwise those with only two classifications would seem overly accurate. In the context of the two

studies that make up this dissertation, there are only two possible classification decisions, mastery or nonmastery, so the percent accurate by chance is 50%.

$$PRE = \frac{(\%\ accurate\ classification - \%\ accurate\ by\ chance)}{(100\% - \%\ accurate\ by\ chance)} \quad (18)$$

4.1.3 ARCH Efficiency (RQ3) Data Collection & Analytic Methods

The third research question addresses the efficiency of testing that results from application of the ARCH approach. The Monte Carlo computer simulations described in the previous section on research question two also collected the test lengths and no-decision rates for SPRT, ARCH pre-calibration, ARCH post-calibration, and EXSPRT based tests that were required to answer research question three.

The initial analytic method proposed for answering research question three was a repeated measures one-way analysis of variance (RM-ANOVA) and subsequent post hoc testing to determine if mean test lengths of ARCH, SPRT, and EXSPRT-R tests were significantly different. A Mauchly's Test of Sphericity would test the homogeneity of covariance (sphericity) assumption on which the RM-ANOVA test is based. Power analysis, using the software application G*Power (Faul et al., 2007) indicated that the 104 examinees in the historical COM test data were sufficient to conduct the planned RM-ANOVA tests able to detect large effect sizes.

However, when the resulting data from the Monte Carlo simulations were examined, the RM-ANOVA analytical method could not be used, as the study data consistently violated the RM-ANOVA assumption of normality. Consequently, the

nonparametric Friedman Test was conducted to see if test length or no-decision rates differed significantly according to the testing algorithm used with post hoc analysis via Wilcoxon signed-rank tests used to determine which specific pairwise comparisons were significant.

4.2 ARCH Evaluation with New IU Plagiarism Test Examinees

The second study of this dissertation, ARCH evaluation with new IU plagiarism test examinees, builds on ARCH operationalization conducted as part of the first study and addresses the three research questions in the context of live testing associated with a new adaptive version of the Indiana University (IU) plagiarism test, which was also created as part of the second study.

Hundreds of universities and colleges direct their students to the plagiarism tutorial and request that their students provide them with the confirmation certificate generated as evidence that they know how to recognize plagiarism. Most of those who take the plagiarism tutorial are early in their post-secondary education. Participants were not offered an incentive for participation in the study. Those who elected to participate and those who declined to participate received exactly the same test with the only difference being the data that were collected behind the scenes. Solicitation of participants took the form of messages with links added to the HTML pages of the current plagiarism tutorial and test that inform users of the option to participate in the study.

With the version of the test in place at the outset of this study, confirmation certificates were generated if an examinee correctly responded to all ten items on the test

associated with the plagiarism tutorial. In each of the ten items, a section of original source material was presented beside a student version that drew on the original source material. Examinees then had to decide if the student version was an example of word-for-word plagiarism, paraphrasing plagiarism, or not an example of plagiarism. Appendix B lists all ten items included in an earlier version of the plagiarism test.

The test associated with the plagiarism tutorial was modified for the purposes of this study and to address two existing issues with the test. The version of the test in place at the outset of this study and included in the plagiarism tutorial always presented the same ten items in the same order, which made cheating easier (e.g. examinees doing the test together or repeatedly to determine correct answers by a process of elimination). Furthermore, after an examinee passed the plagiarism test, they were presented with a confirmation certificate in the form of an HTML page that they could print and hand in to their instructor as evidence that they understood plagiarism. However, examinees could easily create multiple copies of the confirmation certificate by simply printing the HTML page or saving the HTML page digitally and sharing the copies with their peers.

Response data from participants were collected by deploying the new adaptive version of the IU plagiarism test alongside the version of the test in place at the outset of this study. The new adaptive version of the IU plagiarism test was different from the original version in several ways. First, items were now randomly drawn from a large item-bank, thus making it harder for an examinee to cheat.

Second, VL-CCT approaches, specifically SPRT, ARCH, and EXSPRT, were now used to end the test once a mastery or nonmastery classification decision was made by all three approaches. Administration of tests continued until all the algorithms had

made a classification decision or 20 items had been deployed. The 20-item test limit was used to restrict the time it took to take a given test and to reduce item exposure.

Third, the nature of the certificate that confirmed that an examinee understood how to recognize plagiarism was modified to both make it easier for examinees to share the certificate with their instructor and make it much harder to forge certificates. Forth, the implementation of the associated test was updated to improve the user experience on a variety of devices and make the code easier to maintain going forward.

The IU plagiarism test provided an ideal context for addressing the research questions in this study for practical rather than test subject matter specific reasons. The practical need to address issues with cheating on the test also provided an opportunity to make changes to the test that enabled data associated with the research questions to be collected. Furthermore, The high volume of examinees that take the IU plagiarism test meant that data collection would complete quickly. Through my collaboration with the lead creator and maintainer of the IU plagiarism test, Dr. Theodore Frick, I had the ability to make modifications necessary to address the research questions and make other enhancements to the test. The research questions and the ARCH approach are not limited to the specific subject matter of recognizing plagiarism or the context of the IU plagiarism test. Any large-scale computer based test could have been used for this research but the IU plagiarism test was the context that I had access to and required modification to address existing issues.

A large item-bank was created for the new adaptive version of the plagiarism test. In addressing the question of item-bank size for use in Computer Adaptive Testing (CAT), Weiss (1985) states:

> CAT operates most effectively from an item-bank with a large number of items that are highly discriminating and are equally represented across the difficulty-trait level continuum. Satisfactory implementations of CAT have been obtained with item-banks that meet these qualifications with as few as 100 items; however, properly structured item-banks in the range of 150 to 200 items will provide better results. (p. 786)

Based on Weiss's guidelines regarding item-bank size, item creation, pilot testing, and calibration of items continued until the item-bank grew to an appropriately large number of items. The specific number of items is not provided for the sake of test security. Items took the same basic form as those in the current version of the test (see Appendix B). The source material from which the original source text was drawn is the type of material that an undergraduate student would likely make reference to in their class papers.

In order for the proposed ARCH approach to be compared to tradition calibration methods, enough calibration data needed to be collected to satisfy the requirements of the traditional calibration methods. Calibration data takes the form of the number of correct and incorrect responses to each item from examinees that are masters and nonmasters. For the purposes of this study, masters and nonmasters were identified using SPRT set to $P(C|M) = .85$, $P(C|NM) = .50$, false master = false nonmaster = .05. Note two deviations

from SPRT settings used in earlier studies: (1) the false master and false nonmaster error rates were set to .05 instead of .025 and (2) the *P(C|NM)* was set to .50 instead of .60. These deviations reflect the lower stakes of the new adaptive IU plagiarism test and enable the test to be passed with fewer items. Traditional and stricter error rates of .025 and *P(C|NM)* of .60 were piloted and found to result in adaptive tests that were overly difficult for examinees.

Responses from at least 50 masters and 50 nonmasters were used to establish item parameter estimates for use with EXSPRT. Recall that Frick (1992) found that calibration sample sizes of 25 masters and 25 nonmasters was sufficient to enable subsequent testing that applied EXSPRT to make classification decisions within classification error rates established *a priori*. However, this research is based in part on questioning the 25 examinees per classification group guidelines for item calibration and posits that different numbers of examinees may be required to sufficiently calibrate items in different item-banks. Consequently, double the number of responses from nonmasters and masters were used during item calibration for the EXSPRT-R portion of the revised version of the plagiarism test. Collecting response data from 50 nonmasters and 50 masters for every item in the item-bank served to increase the chances that the test was sufficiently calibrated to enable subsequent EXSPRT-R testing within classification error rates established *a priori*. Given that the results of EXSPRT-R testing using items calibrated with a fixed number of masters and nonmasters was the primary means of identifying masters and nonmasters for use in evaluating ARCH accuracy (RQ2), it was important that the conservative 50 responses from each classification group be used to calibrate the items.

The ARCH approach calibrates items until they meet specific calibration criteria (see section *3.6.4 Heuristics*). However, defining the specific ARCH calibration criteria was conducted in the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data. Consequently, it could not be known beforehand precisely how many participants would be involved in calibration of the IU plagiarism test items using the ARCH approach.

The second study achieved a 53% participation rate. During the phase between December 17th, 2013 and January 26th, 2014 when data were collected for the baseline item calibration for use with EXSPRT, test data from nearly 15,000 unique examinees was used to gather 50 responses from nonmasters and 50 responses from masters to every item in pool.

In addition, 5,729 unique examinees volunteered to participate in the second study during the phase between January 26th, 2014 and January 31st, 2014 when data were collected to evaluate ARCH. Unique email addresses provided the method for identifying unique examinees. This brought the total number of participants in the second study to over 20,000 individuals. Examinees took the test over five times on average, with most stopping once they had earned the confirmation certificate. In order to maintain independence of observations, only one test could be used per examinee. Random numbers were associated with every test, and only the test with the smallest random number was selected for inclusion in the study.

A total of 4,641 test administrations were required before 83% of the item bank was calibrated and the EXSPRT component of ARCH could be deployed. This

calibration phase of ARCH took place during a 25-hour period on January 26 and 27, 2014, which illustrates how heavily the IU Plagiarism Test is typically used near the beginnings of college semesters.

Given that power analysis for the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data, revealed that the 104 examinees would be adequate to conduct both tests, the involvement of thousands of participants in the second study indicated that the sample was more than sufficient to enable subsequent statistical testing.

Familywise Error Rate

The familywise error rate for this study was also set to the often used value of $\alpha = .05$. As five hypothesis tests were performed for ARCH pre-calibration and the same five for ARCH post-calibration, the p-value used for testing each of the hypotheses was set to $p = .005$. RQ2 entailed performing statistical analysis on three hypothesis tests: (1) Do overall error rates differ significantly across the two testing algorithms? (2) Do rates of false nonmastery differ significantly across the two testing algorithms? and (3) Do rates of false mastery differ significantly across the two testing algorithms? The remaining two hypothesis tests were associated with RQ3: (3) Do mean test lengths differ significantly across the three testing algorithms? and (4) Do no-decision rates differ significantly across the two testing algorithms? Note that EXSPRT is excluded from the hypothesis tested involving error and no-decisions rates since EXSPRT served as the indicator of the examinee true mastery status.

Post-hoc testing associated with any of the four hypothesis tests listed above – to look for pairwise differences between the four testing algorithms – requires that the p = .005 value be divided again further. Three testing algorithms equated to the following three pairwise comparisons: (1) SPRT and ARCH; (2) SPRT and EXSPRT; (3) ARCH and EXSPRT. Applying the Bonferroni correction resulted in dividing p = .005 by three, which yielded the small value of p = .0017 that was used in post-hoc analyses.

As in study 1, goodness-of-fit tests conducted in study 2 were not included in the overall familywise error rate and individual hypothesis testing p values. Unlike the other statistical tests of significance conducted in study 2, lack of a significant difference between the error rates observed with the ARCH algorithm and the expected *a priori* error rates represents a discovery. Consequently, keeping the p value at .05 is a conservative approach, as the probability of finding a significant result is greater than if p was decreased.

4.2.1 Calibration Sufficiency (RQ1) Addressed by Implication

RQ1 seeks to identify specific ARCH criteria that enable ARCH to reliably predict when an item has been sufficiently calibrated to enable subsequent testing to efficiently make classification within *a priori* error rates. The second study, ARCH evaluation with new IU plagiarism test examinees, unlike the first study, took place in the context of live testing. Therefore, methods of conducting a large number of simulations in order to examine the impact that various ARCH criteria have on classification accuracy were *not* available as they had been in the context of the first study. With the COM test in the first study, examinee responses to all the items in the test were used to establish the mastery or nonmastery classification of the examinee; however, in the live

testing context of the second study, such a means of establishing "true" classifications would not be available, since examinees only responded to a small subset of the test items in the pool. Consequently, the second study *applied* the ARCH criteria established in the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data. If ARCH worked well via use of these established thresholds, and answers to RQ2 and RQ3 were satisfactory in the second study, then by implication the thresholds would be good enough in practice.

4.2.2 ARCH Accuracy (RQ2) Analytic Methods

The second research question addresses the accuracy of testing classification decisions that result from application of the ARCH approach. The classification decisions made using the ARCH approach collected during live testing were compared to classification decisions made using the EXSPRT-R (50) through agreement tables and chi-squared goodness of fit tests. In addition, the nonparametric Wilcoxon signed-rank tests were used to determine if ARCH overall error, false nonmaster, and false master error rates differed significantly from SPRT.

4.2.3 ARCH Efficiency (RQ3) Analytic Methods

The third research question addresses the efficiency of testing that results from application of the ARCH approach. Again, data collected during live testing provided the basis for answering this question.

The nonparametric Friedman Test was used to see if mean test lengths differed significantly according to the testing algorithm used with post hoc analysis via Wilcoxon signed-rank tests, which were used to determine which specific pairwise comparisons

were significant. In addition, the Wilcoxon signed-rank test was used to determine if SPRT and ARCH had significantly different no-decision rates.

CHAPTER V. RESULTS

5.1 ARCH Calibration Sufficiency (RQ1) Results

RQ1 – When is an item sufficiently calibrated? – was answered during the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data. Recall that the first study involved a series of computer simulations using data collected from examinees (n=104) in a previous study (Frick, 1992) who responded to an 85-item test of knowledge of how computers functionally work (COM test).

The examinee's actual results on the original complete 85-item test served as their true mastery status and enabled adaptive test results to be evaluated for accuracy against this benchmark. Of the 104 examinees, 28 qualified as true nonmasters due to a total score that was less than the cutoff of 72.5%, and 76 qualified as true masters by reaching or exceeding the cutoff. Table 15 provides examinee and response statistics by nonmaster and master classification. The number of correct and incorrect responses by nonmasters and masters enabled the probability of a correct response from a nonmaster, $P(C|NM)$, and master, $P(C|M)$, to be calculated for the entire item pool.

Table 14. *COM Test Examinee and Response Statistics By Classification*

Examinees	Nonmaster (NM)				Master (M)			
	#	$\#C_{NM}$	$\#\neg C_{NM}$	*P(C|NM)*	#	$\#C_M$	$\#\neg C_M$	*P(C|M)*
104	28	1338	1042	0.56	76	5656	804	0.88

The process of addressing RQ1 brought challenges whose resolution required deviating from the methods and ARCH criteria initially proposed and a near complete rewrite of the first computer program that was developed to conduct the computer simulations. Ultimately, RQ1 was answered, but the challenges faced and the means of overcoming them also revealed findings of note. The following describes the process that was followed to answer RQ1, the major challenges that occurred, how these challenges were addressed, noteworthy findings associated with overcoming these challenges, and, finally, the answers established for RQ1. The first step taken to address RQ1 was the development of the first version of a web-based computer program to conduct the test simulations.

5.1.1 Version 1 of Web-based Computer Program for Simulations

Version 1 of a web-based computer program was written using several thousand lines of custom HTML and JavaScript code to re-administer simulated adaptive tests using the historical COM test data based on specific input values outlined in more detail below. The output of the simulation was displayed in HTML tables that were then copied to Excel and SPSS for analysis.

The web-based computer program developed allowed calibration, SPRT and EXSPRT settings, and ARCH criteria to be entered as input values prior to running the

simulations (see figure 10). The calibration settings specified the percentage of correct answers on the total test needed to qualify as a master with 72.5% representing the cut-score used in previous COM test studies. The remaining calibration settings dictated how the simulation would execute and when the simulation would terminate.

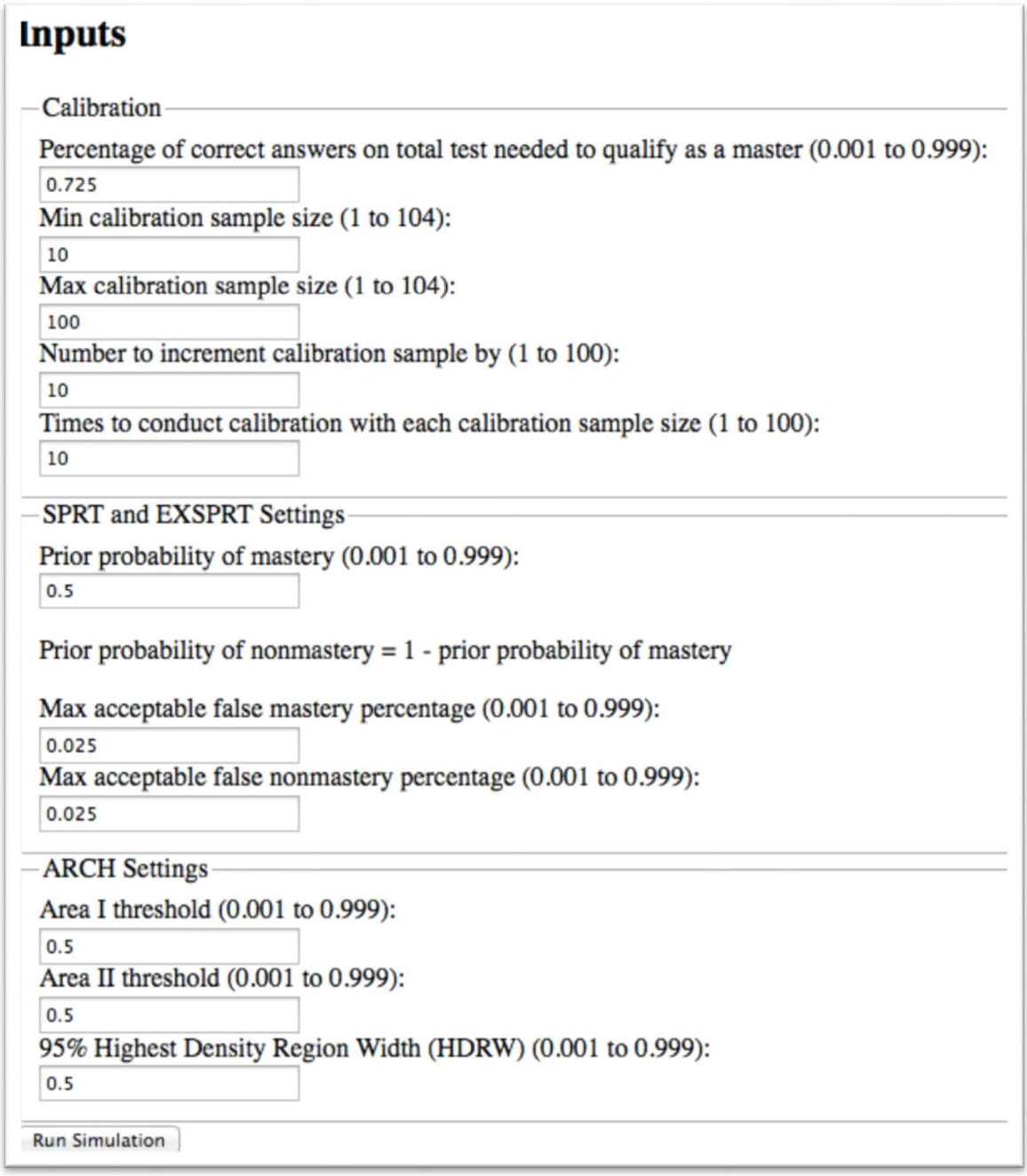

Figure 10. Screenshot of sample input settings for version 1 of the web based Monte Carlo COM Test Simulation program.

For example, the calibration settings values in figure 10 resulted in the following steps occurring during a simulation run.

Simulation Run Steps:

1. Set the calibration sample size to the minimum calibration sample size of 10.
2. Use item responses from randomly selected masters and nonmasters equal to the calibration sample size (e.g. 10 masters and 10 nonmasters) to calibrate items for use with EXSPRT and empirically establish item-bank level probabilities for use with SPRT.
3. Administer a simulated test to each of the 104 examinees with all relevant SPRT and EXSPRT data being output to the associated tables.
4. Repeat steps 2 and 3 the number of times indicated (e.g., 10 rounds).
5. Increment the calibration sample size by the increment value of 10.
6. If the calibration sample size is less than or equal to the maximum calibration sample size value of 100 then go to step 2. Otherwise, end the simulation.

Each of the adaptive testing algorithms also had *a priori* error rates that could be set to specific values. However, the simulations conducted for the purposes of this study used the same values that matched those used in earlier COM test studies. The prior probability of mastery was set to .5 and both the *a priori* false mastery and false nonmastery error rates were set to .025.

In each simulated adaptive test associated with step 3 above, a specific examinee would be randomly administered one of the 85 items with their actual correct or incorrect response to the item being available in the historical data and used as their response in the simulated test. Items would continue to be administered randomly to the same examinee until either all 85 items had been exhausted or all the adaptive testing algorithms had

been able to make a classification decision. The process would repeat with the next examinee and continue in this way until the conditions for the termination of the simulation specified by the calibration inputs had been met.

Each run of the simulations would populate data into five tables: (1) Precision of Item Calibration Estimates; (2) SPRT and EXSPRT Results by Examinee; (3) SPRT Precision of Item Calibration Estimates and Test Metrics By Unique Test; (4) EXSPRT Precision of Item Calibration Estimates and Test Metrics By Unique Test; and (5) Test Metrics By Calibration Sample Size Group. Screenshots of the first few rows of each of the five tables are provided below. It is important to note that these screenshots are not results of the study but are provided to show context on how the simulations operated.

Cal Sample Size	Cal Round	Item ID	Nonmaster (NM)								Master (M)								Beta t-test
			#	s	f	P(C|NM)	P(!C|NM)	Area I	Area II	Beta t-test	#	s	f	P(C|M)	P(!C|M)	Area I	Area II	Beta t-test	
10	1	All	4	176	164	0.518	0.482	0	0	235.327	6	455	55	0.891	0.109	0	0	235.327	144.588
10	1	0	4	1	3	0.333	0.667	0	0.278	5.427	6	6	0	0.875	0.125	0.001	0.359	5.427	3.394
10	1	1	4	2	2	0.5	0.5	0.008	0.442	3.583	6	6	0	0.875	0.125	0.016	0.558	3.583	2.455
10	1	2	4	4	0	0.833	0.167	0.278	0.743	0.499	6	6	0	0.875	0.125	0.512	0.851	0.499	1.119
10	1	3	4	3	1	0.667	0.333	0.059	0.598	2.087	6	6	0	0.875	0.125	0.121	0.785	2.087	1.756
10	1	4	4	1	3	0.333	0.667	0	0.278	5.427	6	6	0	0.875	0.125	0.001	0.359	5.427	3.394
10	1	5	4	4	0	0.833	0.167	0.278	0.743	0.499	6	6	0	0.875	0.125	0.512	0.851	0.499	1.119
10	1	6	4	3	1	0.667	0.333	0.263	0.771	0.78	6	5	1	0.75	0.25	0.367	0.897	0.78	0.86
10	1	7	4	4	0	0.833	0.167	0.278	0.743	0.499	6	6	0	0.875	0.125	0.512	0.851	0.499	1.119
10	1	8	4	2	2	0.5	0.5	0.008	0.442	3.583	6	6	0	0.875	0.125	0.016	0.558	3.583	2.455
10	1	9	4	0	4	0.167	0.833	0	0.115	8.475	6	6	0	0.875	0.125	0	0.21	8.475	5.042
10	1	10	4	3	1	0.667	0.333	0.263	0.771	0.78	6	5	1	0.75	0.25	0.367	0.897	0.78	0.86

Figure 11. Screenshot of sample output of results from estimating precision of item calibrations.

Where:

Cal Sample Size is the calibration sample size being simulated
Cal Round is the calibration round for a given calibration sample size
Item ID is the ID of one of the 85 items being calibrated
Nonmaster (NM) is label for all the nonmaster calibration statistics for the item
Master (NM) is label for all the master calibration statistics for the item
is the number of nonmasters/masters in the calibration sample size
s is the number of nonmaster/master successful/correct responses to the item
f is the number of nonmaster/master failed/incorrect responses to the item
P(C|NM)/P(C|M) is the probability of a correct response from a nonmaster/master
P(!C|NM)/P(!C|M) is the probability of an incorrect response from a nonmaster/master
Area I is the area under masters beta distribution curve between the *P(C|NM)* and the end of the tail
Area II is the area under masters beta distribution curve between the *P(C|NM)* and a specific alpha point
Beta t-tests columns are initial experiments with various versions of the *Beta Difference Index* statistic

Cal Sample Size	Cal Round	Examinee ID	Is Master	SPRT Results					EXSPRT Results					Total Test Total Score
				Correct	False NM	False M	No Dec	Test L	Correct	False NM	False M	No Dec	Test L	
10	1	0	0	1	0	0	0	29	1	0	0	0	32	0.671
10	1	1	1	1	0	0	0	7	1	0	0	0	7	0.824
10	1	2	0	1	0	0	0	6	1	0	0	0	5	0.471
10	1	3	0	1	0	0	0	8	1	0	0	0	11	0.482
10	1	4	0	1	0	0	0	7	1	0	0	0	6	0.282
10	1	5	1	1	0	0	0	60	1	0	0	0	17	0.8
10	1	6	0	1	0	0	0	4	1	0	0	0	19	0.471
10	1	7	0	1	0	0	0	34	0	0	1	0	26	0.706
10	1	8	1	1	0	0	0	11	1	0	0	0	7	0.835
10	1	9	0	1	0	0	0	33	1	0	0	0	46	0.718
10	1	10	0	1	0	0	0	4	1	0	0	0	52	0.588
10	1	11	0	1	0	0	0	3	1	0	0	0	6	0.494
10	1	12	1	1	0	0	0	7	1	0	0	0	8	0.824
10	1	13	0	1	0	0	0	11	1	0	0	0	15	0.541
10	1	14	0	1	0	0	0	7	1	0	0	0	7	0.471
10	1	15	0	1	0	0	0	3	0	0	1	0	19	0.624
10	1	16	0	1	0	0	0	14	1	0	0	0	10	0.576
10	1	17	1	0	1	0	0	44	1	0	0	0	28	0.741

Figure 12. Screenshot of sample results for SPRT and EXSPRT tests during Monte Carlo COM test simulations

Where:

Cal Sample Size is the calibration sample size being simulated
Cal Round is the calibration round for a given calibration sample size
Examinee ID is the ID of one of the 104 examinees who took the COM test
Is Master is the indication of if the examinee is a master based on their total test score
Master (NM) is label for all the master calibration statistics for the item
SPRT Results are the set of results associated with the SPRT based test
EXSPRT Results are the set of results associated with the EXSPRT based test
Correct/False NM/False M/No Dec indicates the accuracy of the test decision
Test Length is the number of items given on the test before a decision was made
Total Test Score is the percent of correct answers the examinee had on the total test

Cal Sample Size	Calibration Round	Precision of Item Calibration Estimates									SPRT Test Metrics						
		Nonmasters				Masters				Beta t-test	PRE	Percent Correct	Percent False NM	Percent False M	Percent No Dec	Test Length	
		#	Area I	Area II	Beta t-test	#	Area I	Area II	Beta t-test							μ	SD
10	1	4	0	0	235.327	6	0	0	235.327	144.588	0.923	96.154	2.885	0.962	0	12.731	9.731
20	1	5	0	-1	264.667	15	0	-1	264.667	199.181	0.788	89.423	9.615	0.962	0	16.731	12.472
30	1	10	0	-1	510.407	20	0	-1	510.407	345.048	0.942	97.115	2.885	0	0	16.452	11.04
40	1	13	0	-1	620.414	27	0	-1	620.414	425.354	0.865	93.269	3.846	2.885	0	17.587	10.538
50	1	13	0	-1	694.819	37	0	-1	694.819	516.172	0.904	95.192	3.846	0.962	0	15.144	11.707
60	1	13	0	-1	650.194	47	0	-1	650.194	510.761	0.904	95.192	3.846	0.962	0	17.49	12.994
70	1	17	0	-1	948.828	53	0	-1	948.828	720.552	0.885	94.231	2.885	2.885	0	12.933	8.969
80	1	21	0	-1	1069.578	59	0	-1	1069.578	793.127	0.846	92.308	6.731	0.962	0	16.356	11.937
90	1	26	0	-1	1285.505	64	0	-1	1285.505	921.775	0.981	99.038	0	0.962	0	18.183	13.151
100	1	27	0	-1	1422.921	73	0	-1	1422.921	1044.494	0.923	96.154	3.846	0	0	15.24	9.383

Figure 13. SPRT precision of item calibration estimates and test metrics by unique calibration round output table screenshot from Monte Carlo COM test simulations

Where:

Cal Sample Size is the calibration sample size being simulated
Cal Round is the calibration round for a given calibration sample size
Precision of Item Calibration Estimates are the set of statistics associated with the item-bank level parameter estimates
Nonmaster (NM) is label for all the nonmaster calibration statistics for the item-bank
Master (NM) is label for all the master calibration statistics for the item-bank
is the number of nonmasters/masters in the calibration sample size
Area I is the area under masters beta distribution curve between the *P(C|NM)* and the end of the tail
Area II is the area under masters beta distribution curve between the *P(C|NM)* and a specific alpha point
Beta t-tests columns are initial experiments with various versions of the *Beta Difference Index* statistic
SPRT Test Metrics are the set of statistics associated with SPRT testing
PRE is the proportion reduction in error achieved by the test
Percent Correct/False NM/False M/No Dec indicates the accuracy of the test decisions
Test Length is the number of items given on the test before a decision was made
μ is the mean test length
SD is the standard deviation associated with the mean test length

Cal Sample Size	Calibration Round	Precision of Item Calibration Estimates															EXSPRT Test Metrics						
		Nonmasters							Masters								PRE	Percent Correct	Percent False NM	Percent False M	Percent No Dec	Test Length	
		#	Area I		Area II		Beta t-test		#	Area I		Area II		Beta t-test									
			μ	SD	μ	SD	μ	SD		μ	SD	μ	SD	μ	SD	μ	SD						μ	SD
10	1	4	0.107	0.134	0.209	0.541	2.95	2.193	6	0.153	0.175	0.2	0.663	2.95	2.193	2.95	2.193	0.923	96.154	0	3.846	0	13.644	11.15
20	1	5	0.059	0.136	0.218	0.359	3.9	2.206	15	0.126	0.198	0.188	0.676	3.9	2.206	3.9	2.206	0.846	92.308	0.962	6.731	0	13.202	8.385
30	1	10	0.04	0.098	0.217	0.322	6.78	3.735	20	0.063	0.12	0.237	0.516	6.78	3.735	6.78	3.735	0.962	98.077	0	1.923	0	14.452	11.152
40	1	13	0.026	0.085	0.194	0.281	8.176	4.168	27	0.041	0.087	0.218	0.5	8.176	4.168	8.176	4.168	0.942	97.087	0	2.913	0.962	16.173	11.567
50	1	13	0.011	0.038	0.176	0.184	9.343	4.425	37	0.03	0.066	0.234	0.471	9.343	4.425	9.343	4.425	0.942	97.115	1.923	0.962	0	12.096	6.765
60	1	13	0.032	0.099	0.216	0.188	8.93	4.938	47	0.068	0.164	0.239	0.533	8.93	4.938	8.93	4.938	0.885	94.231	2.885	2.885	0	13.202	9.923
70	1	17	0.005	0.03	0.142	0.111	12.733	5.74	53	0.016	0.054	0.243	0.391	12.733	5.74	12.733	5.74	0.904	95.192	0.962	3.846	0	10.587	7.491
80	1	21	0.019	0.082	0.184	0.127	14.246	6.857	59	0.029	0.099	0.242	0.367	14.246	6.857	14.246	6.857	0.885	94.231	3.846	1.923	0	12.962	8.431
90	1	26	0.012	0.051	0.177	0.125	16.771	7.92	64	0.022	0.071	0.254	0.318	16.771	7.92	16.771	7.92	0.942	97.115	1.923	0.962	0	13.846	9.825
100	1	27	0.007	0.044	0.146	0.09	18.69	8.435	73	0.014	0.056	0.241	0.286	18.69	8.435	18.69	8.435	0.942	97.115	1.923	0.962	0	12.337	8.036

Figure 14. EXSPRT Precision of Item Calibration Estimates and Test Metrics By Unique Test Output Table Screenshot

Where:

Calibration Sample Size is the calibration sample size being simulated
Calibration Round is the calibration round for a given calibration sample size
Precision of Item Calibration Estimates are the set of statistics associated with the item-bank level parameter estimates
Nonmaster (NM) is label for all the nonmaster calibration statistics for the item-bank
Master (NM) is label for all the master calibration statistics for the item-bank
is the number of nonmasters/masters in the calibration sample size
Area I is the area under masters beta distribution curve between the *P(C|NM)* and the end of the tail
Area II is the area under masters beta distribution curve between the *P(C|NM)* and a specific alpha point
μ is the mean of the associated statistic
SD is the standard deviation associated with mean
Beta t-tests columns are initial experiments with various versions of the *Beta Difference Index* statistic
EXSPRT Test Metrics are the set of statistics associated with EXSPRT testing
PRE is the proportion reduction in error achieved by the test
Percent Correct/False NM/False M/No Dec indicates the accuracy of the test decisions
Test Length is the number of items given on the test before a decision was made

Cal Sample Size	SPRT Results												EXSPRT Results											
	PRE		Percent Correct		Percent False NM		Percent False M		Percent No Dec		Test Length		PRE		Percent Correct		Percent False NM		Percent False M		Percent No Dec		Test Length	
	μ	SD	μ	SD	μ	SD	μ	SD	μ	SD	μ	SD	μ	SD	μ	SD	μ	SD	μ	SD	μ	SD	μ	SD
10	0.923	0	96.154	0	2.885	0	0.962	0	0	0	12.731	9.731	0.923	0	96.154	0	0	0	3.846	0	0	0	13.644	11.15
20	0.788	0	89.423	0	9.615	0	0.962	0	0	0	16.731	12.472	0.846	0	92.308	0	0.962	0	6.731	0	0	0	13.202	8.385
30	0.942	0	97.115	0	2.885	0	0	0	0	0	16.452	11.04	0.962	0	98.077	0	0	0	1.923	0	0	0	14.452	11.152
40	0.865	0	93.269	0	3.846	0	2.885	0	0	0	17.587	10.538	0.942	0	97.087	0	0	0	2.913	0	0.962	0	16.173	11.567
50	0.904	0	95.192	0	3.846	0	0.962	0	0	0	15.144	11.707	0.942	0	97.115	0	1.923	0	0.962	0	0	0	12.096	6.765
60	0.904	0	95.192	0	3.846	0	0.962	0	0	0	17.49	12.994	0.885	0	94.231	0	2.885	0	2.885	0	0	0	13.202	9.923
70	0.885	0	94.231	0	2.885	0	2.885	0	0	0	12.933	8.969	0.904	0	95.192	0	0.962	0	3.846	0	0	0	10.587	7.491
80	0.846	0	92.308	0	6.731	0	0.962	0	0	0	16.356	11.937	0.885	0	94.231	0	3.846	0	1.923	0	0	0	12.962	8.431
90	0.981	0	99.038	0	0	0	0.962	0	0	0	18.183	13.151	0.942	0	97.115	0	1.923	0	0.962	0	0	0	13.846	9.825
100	0.923	0	96.154	0	3.846	0	0	0	0	0	15.24	9.383	0.942	0	97.115	0	1.923	0	0.962	0	0	0	12.337	8.036

Figure 15. Test Metrics By Calibration Sample Size Group Output Table Screenshot

Where:

Cal Sample Size is the calibration sample size being simulated
SPRT/EXSPRT Results are the set of statistics associated with SPRT/EXSPRT test simulations
PRE is the proportion reduction in error achieved by the test
Percent Correct/False NM/False M/No Dec indicates the accuracy of the test decisions
Test Length is the number of items given on the test before a decision was made
μ is the mean of the associated statistic
SD is the standard deviation associated with mean but was not calculated in this version of the simulation

5.1.2 High Error Rate with Empirically Established Item-bank Level Probabilities

The first challenge experienced while answering RQ1 resulted from the use of empirically established item-bank level probabilities with SPRT instead of the set item-bank level probability values used in previous studies. It was found that the use of empirically established item-bank level probabilities with SPRT resulted in false nonmaster error rates higher than those established *a priori*. This was a problem for two reasons. First, the SPRT algorithm is a key component of ARCH, and issues with SPRT would also represent issues with ARCH. Second, RQ2 involves the comparison of ARCH to SPRT and problems with SPRT would limit the value of this comparison. The SPRT challenge was overcome by setting item-bank level probabilities to values used in previous COM test data-based studies rather than applying an empirical approach.

Item-bank level probabilities of a correct response from each classification group (equations 2 and 4) were established empirically using the total number of correct and incorrect responses from nonmasters ($\#C_{NM}$ and $\#\neg C_{NM}$) and masters ($\#C_M$ and $\#\neg C_M$). A probability of a correct response given nonmastery, *P(C|NM)*, of 0.56, was established empirically by using all available response data from true nonmasters. A probability of a correct response given mastery, *P(C|M)*, of 0.88 was also established empirically through response data from all true masters. The index of discrimination (equation 15) for the item-bank level probabilities established empirically was 0.32.

Item-bank level probabilities used with SPRT have not typically been established empirically in previous studies, with the exception Welch and Frick's (1993, p. 57) use of empirically derived values for use with SPRT. In the original research that used COM test data, "the SPRT parameters were set *a priori* as follows: mastery level = .85, non-mastery level = .60, $\alpha = \beta = .025$" (Frick, 1989, p. 102) instead of establishing the values empirically.

The values .85 and .60 were selected to reflect widely used letter grade cutoffs with .725 representing the value between these two cutoffs. Using the set values from the Frick (1989) study, the index of discrimination for the item-bank level probabilities (equation 15) was 0.25, which is 0.07 or 21.9% smaller than the index of discrimination calculated using empirically established probabilities. In other words, the set SPRT parameters were substantially less discriminating than the SPRT parameters established empirically.

I initially thought that empirically based item-bank level probabilities would be the most appropriate to use in simulations involving both SPRT and ARCH. Since these values are based on actual response data, I expected them to lead to optimal SPRT performance. However, repeated simulations of SPRT with COM test data using $P(C|M) = 0.88$, $P(C|NM) = 0.56$, $\alpha = \beta = .025$ consistently yielded false nonmastery error rates that exceeded the .025 rate established *a priori*. Recall that a false nonmastery error occurs when an examinee is classified as a nonmaster when they are, in fact, a master (according to their total test score).

For example, examinee response data from 104 examinees was used to simulate 2,080 SPRT tests calibrated empirically using all the available response data to establish item-bank level probabilities of $P(C|M) = 0.88$, $P(C|NM) = 0.56$, $\alpha = \beta = .025$. The simulation involved each examinee being administered a SPRT based test twenty times. Since SPRT randomly selects items, the chances of two SPRT test administrations being identical is unlikely.

Table 15 provides the error rates by algorithm and includes both the empirically calibrated SPRT and manually calibrated SPRT using set parameters to match earlier SPRT studies based on COM test data. Out of the 2,080 simulated test administrations, SPRT calibrated *empirically* made 2,073 decisions with 67 of those decisions (3.23%) being false nonmastery decisions. A nonmastery error rate of 3.23% is above the *a priori* false

nonmastery rate of 2.5% by 0.73%. The mean test length of tests that applied SPRT calibrated empirically was 16.61 items (*SD* = 12.28).

Table 15. *SPRT Decision Error Rates By Method of Setting Item-bank Level Probabilities*

Item-bank Probability Approach	False Nonmastery Errors		False Mastery Errors		Total Errors		Total Nonmastery & Mastery Decisions
	n	*%*	*n*	*%*	*n*	*%*	*N*
Empirical	67	3.23%	29	1.40%	96	4.63%	2,073
Manual	42	2.11%	28	1.41%	70	3.52%	1,990

Using the same approach, examinee response data from 104 examinees was used to simulate 2,080 SPRT tests using item-bank level probabilities *manually* set to $P(C|M) = 0.85$, $P(C|NM) = 0.60$, $\alpha = \beta = .025$ that are consistent with earlier studies (Frick, 1989). Out of the 2,080 simulated test administrations, SPRT made 1,990 decisions (83 fewer than the empirically calibrated SPRT) with 42 of those decisions (2.11%) being false nonmastery decisions – well below the *a priori* false nonmastery rate of 2.5%. The mean test length of the tests that applied SPRT with item-bank probabilities set manually was 21.97 items (*SD* = 16.37), which is 5.36 items (32.3%) longer than the results obtained with SPRT using item-bank probabilities set empirically. On hindsight, this should not be surprising, since the SPRT requires more items to reach a decision when Wald's zone of uncertainty is smaller ([.85 - .60 = .25] is less than [.88 - .56 = .32]), when using the same *a priori* error rates (see Frick, 1989).

The differences in the results between the SPRT algorithms calibrated empirically and using values set manually are outside the scope of this study but warrant further investigation. Nevertheless, results from the analysis above show that: (1) it cannot be assumed that SPRT will always make classification decisions within error rates established *a*

priori; (2) choice of item-bank level probabilities impacts SPRT error rates; and (3) the empirically established item-bank level probabilities for the COM test data had a higher index of discrimination, shorter average SPRT test lengths, and higher false mastery error rates than associated manually set item-bank level probabilities.

Given that SPRT using set values used in earlier studies resulted in decision error rates within rates established *a priori*, I decided to proceed with SPRT using the manually set values and abandon empirically established item-bank level probabilities for use with SPRT for the remainder of the Monte Carlo studies.

5.1.3 Initial Calibration Statistics Problematic

The first use of computer simulations in the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data, required programming and then testing the effectiveness of the calibration measures proposed for use with the heuristics that determine when an item is sufficiently calibrated. In some cases, the measures originally proposed relied on intensive calculations to establish the area under specific areas of the unique beta distributions associated with $P(C_i|M)$ and $P(C_i|N)$ *every time* new calibration data were collected. This proved to both be programmatically complex and contribute to slow system performance that would likely result in unacceptable delays during live testing when the test is administered online via a Web server—examinees could be waiting excessively long times for the next test question, especially when Web servers are under high demand. In other cases, the proposed measures only functioned acceptably within limited circumstances.

More practical and robust alternatives were found or established that reduce calculation complexity, correlate strongly with the originally proposed measures, and align with established statistical approaches. The following will: (1) describe the rationale for

moving away from both the 95% Highest Density Region Width calculation and calculations of the overlapping areas of $P(C_i|M)$ and $P(C_i|N)$ and (2) explain the associated alternatives that were identified or established.

95% Highest Density Region Width Replaced With Standard Deviation

Recall that the 95% highest density region width (HDRW) was proposed as a measure of the precision of the $P(C_i|M)$ and $P(C_i|N)$ estimated beta means, since the 95% HDRW would get smaller as estimated values became more precise. However, it was found that calculating the 95% HDRW was much more involved than expected.

No equation or set of equations was found for calculating HDRW. Instead, establishing the HDRW of a given beta density function is an optimization problem that involves finding the narrowest region under a given beta density function that equals a given probability (e.g. 95%). Finding such a region is straightforward. The challenge is finding the narrowest such region from among the possible areas. Textbooks that address the concept of a beta density function highest density region (HDR) often provide tables in an appendix that enable one to look up the HDR for a limited set of specific functions rather than providing a method for coming up with HDRs.

An approach for determining the 95% HDRW was programmed but it involved performing hundreds or thousands of calculations and dramatically decreased the speed of the simulations to unacceptable levels. It was also clear that the computationally intensive calculation of the 95% HDRW could pose significant problems during the second study, ARCH evaluation with new IU plagiarism test examinees, when the 95% HDRW would need to be calculated during real-time execution of the test associated with the Indiana University

plagiarism tutorial. Consequently, alternative approaches for determining the precision of the $P(C_i|M)$ and $P(C_i|N)$ were explored.

I subsequently determined empirically that the standard deviation associated with its beta density function proved to be an excellent alternative to the 95% HDRW, since it provides a measure of the precision beta mean via a single straightforward calculation. The following explains the equations used to calculate the beta *SD* and the relationship found between the beta *SD* and the 95% HDRW. The equations for the beta mean, variance, and standard deviation are presented below in equations 19, 20, and 21.

$$Beta\ Mean = E[X] = \frac{\alpha}{\alpha + \beta} \tag{19}$$

$$Beta\ Variance = var[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{20}$$

$$Beta\ Standard\ Deviation = SD[X] = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}} \tag{21}$$

As in the 95% HDRW, the beta *SD* also gets smaller as the estimate of the beta mean becomes more precise. Unlike the 95% HDRW, the beta *SD* is straightforward to calculate, which is advantageous both from a programming perspective and for explanatory purposes.

When applying beta distribution equations to adaptive testing, Frick (1992) used a slightly different beta mean equation, based on Schmidt (1969) to ensure that the beta mean would: (1) always be a positive non-zero value and (2) equal 0.5 when both parameters are zero.

$$\alpha = s + 1 \tag{22}$$

$$\beta = f + 1 \tag{23}$$

$$s + f = N \tag{24}$$

Equations 25, 26, and 27 are the resulting beta equations derived by substituting equations 22, 23, and 24 into equations 19, 20, and 21.

$$Beta\ Mean = \frac{s+1}{(s+1)+(f+1)} = \frac{s+1}{N+2} \tag{25}$$

$$Beta\ Variance = \frac{(s+1)(f+1)}{[(s+1)+(f+1)]^2[(s+1)+(f+1)+1]} \tag{26}$$

$$Beta\ Variance = \frac{(s+1)(f+1)}{(N+2)^2(N+3)}$$

$$Beta\ Standard\ Deviation = \sqrt{\frac{(s+1)(f+1)}{(N+2)^2(N+3)}} \tag{27}$$

In subsequent evaluation of the relationship between the 95% HDRW and the beta *SD*, I found a near-perfect correlation between the two. Using the Highest Density Region table provided in Schmidt (1969, p. 378), 142 95% HDRWs were calculated for *s* and *f* values ranging from 0 to 50. Beta *SD* values were calculated based on the same *s* and *f* values. The Pearson correlation between the 95% HDRW and the associated beta *SD* values was .987—a near-perfect relationship. This finding supports the use of the beta *SD* as a replacement for the 95% HDRW.

In the context of evaluating the precision the beta mean of $P(C_i|M)$ or $P(C_i|N)$ to determine when enough data has been collected, a minimum value for the beta *SD* would need to be established. In other words, at what point has the beta *SD* of a beta distribution associated with an item decreased enough to indicate that the associated beta mean for that item is sufficiently precise?

Overlapping Areas of $P(C_i|M)$ and $P(C_i|N)$ Replaced with Beta Difference Index. Originally, the intention was to use a calculation of the overlapping area associated with the beta distributions for $P(C_i|M)$ and $P(C_i|N)$ to calculate the probability that $P(C_i|M)$ is larger or smaller than $P(C_i|N)$. For example, the area under the $P(C_i|M)$ beta distribution to the left of the beta mean of the $P(C_i|N)$ beta distribution represents the probability that $P(C_i|M)$ is less than the beta mean of the $P(C_i|N)$ and can be calculated by using Simpson's rule to integrate under specific portions of the beta distribution curve.

There were two problems with using the overlapping area between $P(C_i|M)$ and $P(C_i|N)$ as a measure, that became clear during computer simulations. First, while considerably simpler than the process for determining the HDRW, integration using Simpson's rule to calculate overlapping areas is still computationally intensive and contributes to slow performance. Second and most problematic was the issue uncovered during computer simulations that overlapping areas of $P(C_i|M)$ and $P(C_i|N)$ were extremely small for many $P(C_i|M)$ and $P(C_i|N)$ combinations, which hindered the utility of using the size of an overlapping area as a measure of the difference between $P(C_i|M)$ and $P(C_i|N)$.

The overlapping area between $P(C_i|M)$ and $P(C_i|N)$ can vary from zero to one, where zero indicates there is no overlap, and one indicates there is a perfect overlap. As $P(C_i|M)$

and $P(C_i|N)$ become more different, the overlapping area gets closer and closer to zero. However, the overlapping area gets very close to zero more quickly than expected, which limits the utility of the overlapping approach to detect greater differences between $P(C_i|M)$ and $P(C_i|N)$ beyond a point. The problem with using the overlapping area to measure differences between $P(C_i|M)$ and $P(C_i|N)$ is analogous to using a thermometer that does not go below zero to measure how cold it is – after a point, the reading from the instrument remains stuck at zero despite noticeable changes in what you want to measure.

For example, consider the following case using only 14 responses from both masters and nonmasters to item *i*. The $P(C_i|M)$ beta distribution curve associated with 9 correct responses and 5 incorrect responses to item *i* from true masters has a mean of .625. The $P(C_i|N)$ beta distribution curve associated with 3 correct responses and 11 incorrect responses to item *i* from true masters has a mean of .25, which makes the index of discrimination (difference between the means) for the item .375. The probability that $P(C_i|M)$ is less than the $P(C_i|N)$ mean of .25 can be calculated by finding the area under the $P(C_i|M)$ beta distribution curve between 0 and .25. Performing this calculation yields an area just smaller than .001, which means that: (1) there is less than a one-in-a-thousand probability that $P(C_i|M)$ is less than 0.25 and (2) the lowest end of the range for the area under this portion of the $P(C_i|M)$ curve, zero, has very nearly been reached after only 14 responses from both masters and nonmasters.

The reality that the overlapping areas of $P(C_i|M)$ and $P(C_i|N)$ nearly disappear after so few responses have been gathered (e.g., only 14 responses from both masters and nonmasters) limits the utility of using the overlapping area as a measure of the difference between $P(C_i|M)$ and $P(C_i|N)$ beta distributions. Items with an index of discrimination much

greater than .375 and based on responses from many more masters and nonmasters than the case described above would have beta distributions with overlapping areas that are similarly miniscule to the case above, despite the fact that we can be much more confident that $P(C_i|M)$ is different from $P(C_i|N)$. A different measure was required that would reveal differences across a greater range of mastery and nonmastery beta distribution curves.

Therefore, I sought a measure of the difference between $P(C_i|M)$ and $P(C_i|N)$ beta distributions that would address both of the problems associated with using the overlapping areas of the beta distributions. The results of the measure needed to reflect the fact that confidence that $P(C_i|M)$ and $P(C_i|N)$ beta distributions are different increases as (1) the differences between $P(C_i|M)$ and $P(C_i|N)$ means grows or (2) variance associated with the $P(C_i|M)$ and $P(C_i|N)$ beta distributions decreases. Furthermore, the measure needed to be straightforward to calculate so that it would not contribute to slow Web server performance during real-time test administrations.

In the same way that the beta *SD* was ultimately used in place of the 95% HDRW, I proposed the *Beta Difference Index* as the replacement for using overlapping areas of $P(C_i|M)$ and $P(C_i|N)$. The *Beta Difference Index* provides a computationally straightforward way of measuring the confidence that $P(C_i|M)$ and $P(C_i|N)$ beta distributions are different, where a larger result indicates more confidence that there is a difference.

Unlike the beta *SD* equation, which is an established equation, the *Beta Difference Index* equation is, to the best of the author's knowledge, new. I derived the *Beta Difference Index* equation by substituting beta mean and beta variance equations into Welch's *t*-test equation. While Welch's *t*-test assumes that means and variances are being drawn from

normally distributed populations, the use of beta mean and beta variance equations in the proposed *Beta Difference Index* suggests that assumptions of normality need not apply to the *Beta Difference Index*.

Welch's *t*-test in equation 28 provides a measure of the confidence that two means are different when equal variances cannot be assumed. The numerator provides a measure of variance *between* the two groups via the difference between the associated means. The denominator provides a measure of variance *within* groups via a calculation using associated variances and sample sizes.

Applying equation 25 for the beta mean and equation 26 for the beta variance leads to a proposed *Beta Difference Index* equation (29).

$$Welch's\ t\ test = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{28}$$

$$Beta\ t\ test = \frac{E[X_1] - E[X_2]}{\sqrt{\frac{var[X_1]}{N_1} + \frac{var[X_2]}{N_2}}} \tag{29}$$

Finally, equation 30 is the resulting proposed *Beta Difference Index* derived by substituting equations 22, 23, and 24 into equation 29.

$$Beta\ t\ test = \frac{\left(\frac{s_1 + 1}{N_1 + 2}\right) - \left(\frac{s_2 + 1}{N_2 + 2}\right)}{\sqrt{\frac{(s_1 + 1)(f_1 + 1)}{(N_1 + 2)^2(N_1 + 3)}/N_1 + \frac{(s_2 + 1)(f_2 + 1)}{(N_2 + 2)^2(N_2 + 3)}/N_2}} \tag{30}$$

The *Beta Difference Index* has several desirable characteristics. While equation 30 above is not simple, it is straightforward to compute and would not contribute to slow system performance in either simulated or live testing. Second and more important, the *Beta Difference Index* is an effective measure of the difference between any two beta distributions rather than being limited to a specific range as was the case with using overlapping areas of beta distribution curves. The *Beta Difference Index* approaches zero as the difference between the two beta distributions decreases but only reaches zero if the beta means are identical. The *Beta Difference Index* continues to increase as confidence that the two beta distributions are different increases either through: (1) a greater difference between beta means in the numerator; or (2) reduction in the variance of either beta distribution in the denominator.

5.1.4 Proposed Item-Level Calibration Criteria Not Sufficient

I determined during computer simulations that the two new item-level criteria (min beta *SD* and the max *Beta Difference Index*) alone were not sufficient to yield test decision accuracy rates within expected levels. I found that, through repeated simulations with COM test data, the most discriminating items would be the first ones to be approved for use with EXSPRT-R as they reached max *Beta Difference Index* criteria. Consequently, early EXSPRT-R decisions would be made exclusively with the most discriminating items and, as such, were prone to making classification errors at rates higher than those set *a priori*.

I added a third calibration measure, percent items approved, to ensure that items with a broader range of discriminations would be available when EXSPRT-R was first deployed. Percent items approved ranges from 0% to 100% and refers to the percentage of items in the pool that had met the two item-level approval criteria: (1) the minimum beta *SD* had been

reached for both the mastery and nonmastery beta distributions of an item and (2) the maximum *Beta Difference Index* had been reached for the item. Unlike the item-level calibration measures, the measure for the percent of items approved refers to the entire item pool, so it is a pool-level calibration measure. For example, setting percent of items approved to 50% means that items that meet both item-level approval criteria would not be used until half of all items in the pool had also met both item-level approval criteria.

I thus needed to determine the specific combination of settings for the single pool-level approval criteria, percent items approved, and the two item-level criteria (beta *SD* and *Beta Difference Index*) that would lead to classification decision making within expected error rates.

5.1.5 Version 2 of Web-based Computer Program for Monte Carlo Simulations

With the discovery that an item-level criterion alone would not be sufficient to enable ARCH to calibrate items in a way that supported accurate and efficient testing, I developed a second version of the web-based computer program to conduct the Monte Carlo simulations using the historical COM test data. While this new version did draw on some of the existing code developed in the first version of the program, I created new software (Version 2) to carry this out.

The second version used the same approach of presenting an input form that collected information on how the simulations would proceed and HTML tables to summarize overall results of the simulation and to generate specific test results down to individual responses to items on specific tests associated with a given examinee that are similar to the tables already presented for the first version of the simulation. The second version of the simulation enabled repeated Monte Carlo simulations to be conducted on the ARCH approach with various

settings in order to establish the specific set of ARCH calibration criteria threshold values that led to accurate and efficient testing.

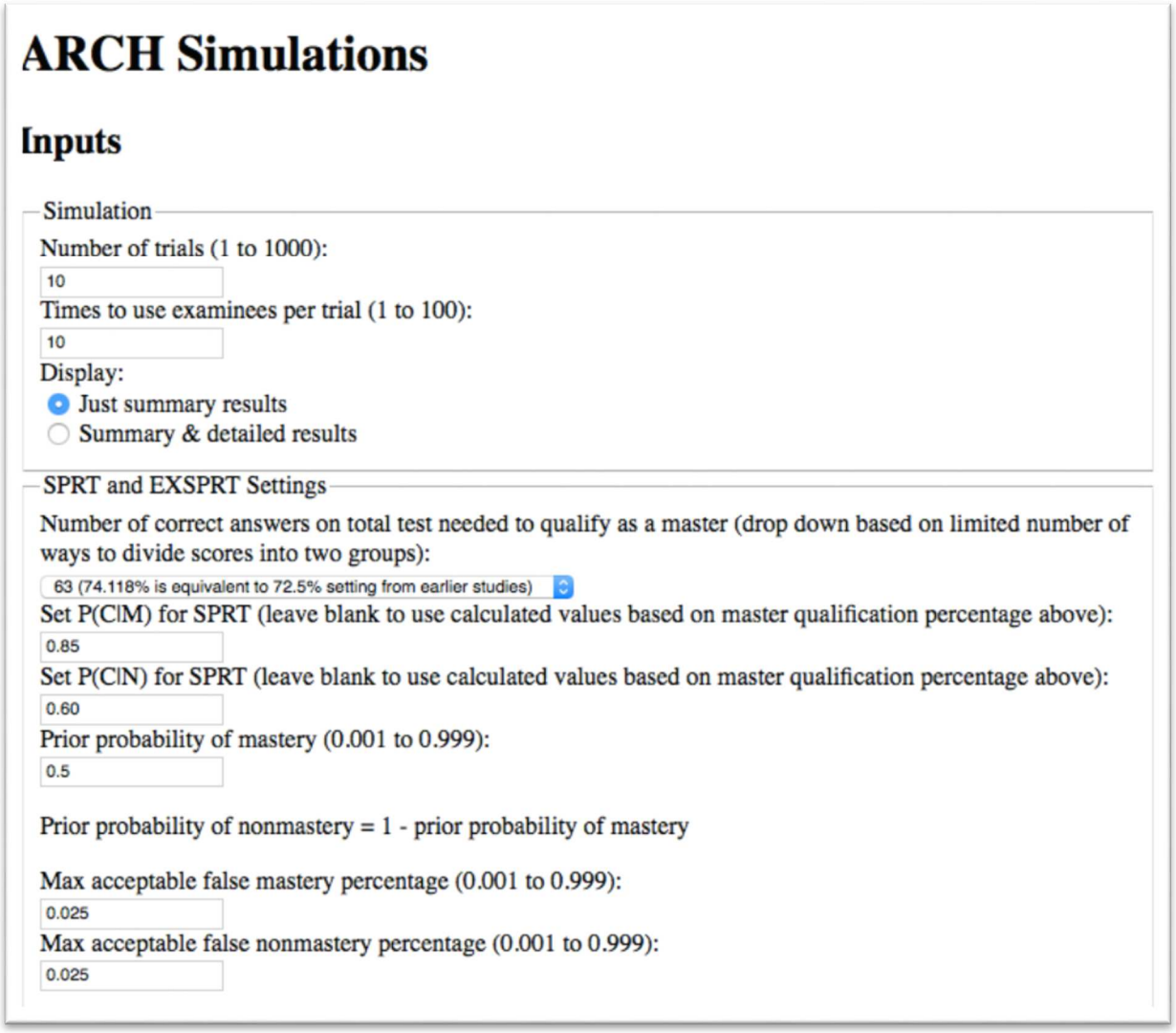

Figure 16. Screenshot of sample input settings for Version 2 of the web-based Monte Carlo COM Test Simulation program.

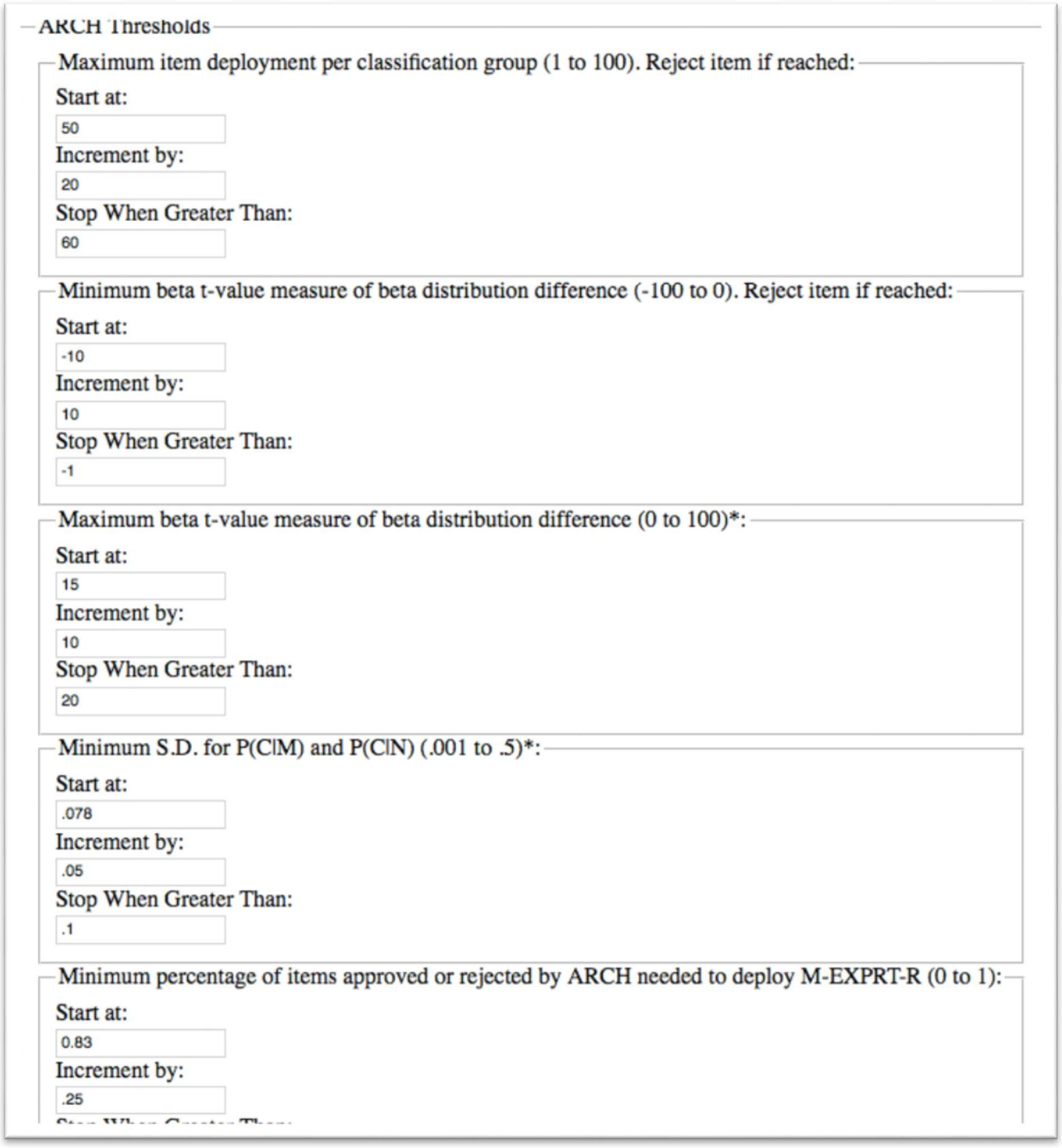

Figure 17. Screenshot of sample input ARCH settings for Version 2 of the web based Monte Carlo COM Test Simulation program.

5.1.6 Establishment of Calibration Statistic Thresholds

The process by which the criteria threshold were chosen involved: (1) establishing lenient and strict threshold bounds for each of the calibration criteria; (2) determining

conservative threshold values for each of the calibration criteria that, when used in isolation, led to sufficiently calibrated items; and (3) systematically adjusting the threshold values for the three calibration criteria to establish a set of threshold values that strike a good balance between ensuring subsequent testing accuracy while hastening the deployment of M-EXSPRT-R to improve test efficiency. Overly conservative calibration thresholds would promote classification accuracy at the expense of test efficiency. Overly lenient calibration thresholds would have the opposite effect – very efficient but unacceptably inaccurate tests. Each of the three steps presented above will be discussed in turn below in the context of each of the three calibration criteria.

Table 16. *Calibration Statistic Bounds and Thresholds*

Statistic	Associated Question	Lenient Bound	Strict Bound	Threshold Value Established
Min Beta *SD*	$P(C_i\|M)$ and $P(C_i\|N)$ estimates precise?	.22	.069	.078
Max *Beta Difference Index*	Are $P(C_i\|M)$ and $P(C_i\|N)$ estimates different?	0	29	15
Max % Items Approved	Are sufficient items approved?	0%	100%	83%

The minimum beta distribution standard deviation (min beta *SD*) provides a measure of the precision of associated beta distribution means $P(C_i|M)$ and $P(C_i|N)$. As the beta standard deviation decreases, the precision of the associated beta distribution mean increases.

The value .22 was used as the lenient bound value of the min beta *SD* since it is associated with the beta distribution formed by one correct response and one incorrect

response. Collecting only two responses seemed unlikely to be sufficient for item calibration. Therefore, .22 provided a reasonable lenient bound for the minimum beta distribution.

The strict bound of .069 for the min beta *SD* was formed based on beta distributions developed by collecting 50 responses where half are correct and half are incorrect. Fifty responses represent a doubling of Frick's (1992) finding that 25 responses by examinees from each classification group to each item led to accurate classification decisions. An equal ratio of correct and incorrect responses corresponds to the largest beta distribution standard deviations for a given calibration sample size and was used to ensure items with this pattern of correct and incorrect responses could possibly make the cutoff.

The lenient bound of 0 corresponds to cases when there is no difference between $P(C_i|M)$ and $P(C_i|N)$. The upper bound of 29 corresponds to the case when the difference between $P(C_i|M)$ and $P(C_i|N)$ is greatest after collecting 50 responses from masters and nonmasters. The maximum beta distribution *t*-test equals 29 after all 50 masters respond correctly to an item and all 50 nonmasters respond incorrectly to the same item.

Table 17. *Example of intermediate steps towards ARCH criteria*

Min Beta *SD*	Max *Beta Difference Index*	Max % Items Approved	False Mastery Rate	False Nonmastery Rate	M-EXSPRT-R Test Length	
					μ	SD
.079	11	83%	1.089%	2.208%	19.228	14.864
.079	12	83%	1.080%	2.345%	19.856	15.095
.079	13	83%	1.035%	2.236%	19.400	14.903
.079	14	83%	1.209%	2.134%	19.500	14.971
.079	15	83%	0.974%	2.455%	19.939	15.222

Finally, the percent of items approved provides an item-pool level setting for ensuring that sufficient calibration data have been collected to enable M-EXSPRT-R to make classification decisions within *a priori* established error rates. The values of 0% and 100% provide the lenient and strict bounds, respectively, for the percent-of-items-approved setting.

Table 17 presents an example of a few of the statistics that were generated using the simulations where the beta *SD*, the *Beta Difference Index*, and the percent items approved were set to specific values. Results of the simulation enabled false mastery rates, false nonmastery rates, and mean test lengths to be compared for different combinations of calibration criteria threshold values. Since random selection of items occurred with every test that made up the simulations, repeatedly using the same thresholds would be unlikely to produce the same results. I continued the simulations until I arrived at a set of calibration threshold values that consistently performed better than other sets of values. However, repeating this simulation process would not necessarily yield the exact same set of values due to the random nature of the item selection methods used in ARCH. Further, it is not clear that using a different data set would yield the same values. Due to the unpredictable nature of testing approaches that use random item selection, the final set of calibration threshold values arrived at should not be considered as *the* only possible combination that will lead to accurate and efficient testing.

5.1.7 RQ1 Results for Monte Carlo ARCH Operationalization and Evaluation with Historical COM Test Data

The ARCH statistic thresholds established through Monte Carlo simulations using historical COM test data and identified in Table 16 consistently led to accurate and efficient testing using the ARCH approach. Therefore, I concluded that an item was sufficiently

calibrated for use with M-EXSPRT-R when (1) the minimum beta distribution for both $P(C_i|M)$ and $P(C_i|N)$ had reached .078, (2) the maximum beta distribution *t*-test of the difference between $P(C_i|M)$ and $P(C_i|N)$ had reached 15, and (3) the percentage of items approved had reached 83%.

5.2 ARCH Accuracy (RQ2) Results

The research question – How accurate is ARCH in comparison to traditionally calibrated SPRT and EXSPRT? – was answered in both the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data, and in the second study, ARCH evaluation with new IU plagiarism test examinees. In both studies, the accuracy of ARCH was examined both pre-calibration and post-calibration. ARCH pre-calibration refers to the period where ARCH testing mimics the SPRT testing algorithm and gathers calibration data on test items. Once sufficient item calibration data had been collected, the period of ARCH post-calibration began. ARCH post-calibration placed the SPRT method and a modified version of EXPSRT in a race to make a classification decision—with the test ending when one of the two can make a decision.

I used similar analytic methods to evaluate the accuracy of ARCH pre-calibration and ARCH post-calibration in both studies—with a few differences that are outlined in the respective sections below. In each section, I first will present the results of goodness-of-fit tests that evaluated whether or not ARCH accuracy results differed significantly from the *a priori* error rates. I will then present the results of nonparametric statistical tests, and note significant differences among the testing algorithms for false nonmastery and false mastery error rates.

5.2.1 RQ2 Results for Monte Carlo ARCH Evaluation with Historical COM Test Data

Chi-Squared Tests

One hundred Chi-Squared tests were conducted on the 50 sets of ARCH pre-calibration tests of 104 examinees and the 50 sets of ARCH post-calibration tests of 104 examinees. Unlike mean error rates and mean test lengths based on 50 tests for each examinee that can be statistically compared across the 104 examinees, the Chi-Squared test depends on each examinee being associated with a single mutually exclusive category. Consequently, a single Chi-Squared test cannot be conducted on all the data generated through 50 sets of 104 simulated tests given four ARCH pre-calibration and ARCH post-calibration tests. Rather than reporting on the results of all 100 chi-squared tests associated with the 50 ARCH pre-calibration samples and the 50 ARCH post-calibration samples, only the results for noteworthy sets and overall findings are presented.

None of the 50 sets of ARCH pre-calibration tests had error rates that deviated significantly ($p \leq .05$) from error rates established *a priori* according to chi-squared goodness-of-fit tests; however, two of the 50 sets of ARCH post-calibration did. Set 37 included seven false nonmaster decisions, one false master decision, and 96 correct classification decisions which resulted in a chi-squared test that indicated a significant departure from *a priori* error rates, $\chi^2=(2, N = 104) = 8.51$, $p = .02$. Set 39 had seven false nonmaster decisions, one false master decision, and 96 correct classification decisions which resulted in a chi-squared test that indicates a significant departure from *a priori* error rates,

χ^2=(2, N = 104) = 7.05, p = .04. All remaining 48 sets for ARCH post-calibration did not reveal significant deviations from error rates established *a priori* according to chi-squared testing. With the Type I error rate of p <= .05, 2.5 of the 50 tests (0.05 x 50) would be expected to be significant by chance alone (attributable to sampling error when the null hypothesis is true in the theoretical chi-squared distribution). Two such findings here are in line with such expectations (2 compared with 2.5).

Overall Error Rate and PRE Descriptive Statistics

The overall error rate and proportion of reduction in error (PRE) are measures that combine the false nonmaster error rate and the false master error rates. Descriptive statistics are provided for the overall error rate in Table 18 and the PRE in Table 19 for each of the four testing algorithms. The statistics are the result of combining the results for all the tests associated with a given examinee into a single examinee-specific value and then computing descriptive statistics for the group of 104 examinees. Consider the 50 simulated ARCH pre-calibration tests conducted with examinee 26, there were no false nonmaster decisions, one false master decisions, and 49 correct classification decisions, which results in an overall error rate of 2% and a PRE of .96. With overall error rates and PRE values available for the remaining examinees, descriptive statistics and tests of normality were calculated.

Table 18. *COM Test Simulations Overall Error Rate Descriptive Statistics and Test of Normality*

					Test of Normality		
Algorithm	*M*	*SD*	Skewness	Kurtosis	Shapiro-Wilk	df	*p*
SPRT	2.77	6.40	3.15	10.72	.50	104	< .001
ARCH Pre-Cal	3.35	8.05	3.29	12.13	.49	104	< .001
ARCH Post-Cal	4.58	9.96	2.98	9.70	.53	104	< .001
EXSPRT	3.29	7.54	3.26	11.43	.50	104	< .001

As shown in table 18, the mean overall error rates for all the testing algorithms were below 5%, which is the combined *a priori* error rate of the false nonmasters error rate (2.5%) and the false master error rate (2.5%). The extreme skewness and kurtosis values reflect that most examinees had very low or zero overall error rates, while those examinees near the cut-score of 72.5% on the total test had higher overall error rates.

Examinees near the cut-score had high frequencies of false nonmaster and false master errors. In fact, the eight examinees with total test scores of 74.12%, 76.47%, or 77.65%, representing less than 8 percent of the examinees, accounted for over 60% of the false nonmaster decisions across all the testing algorithms in the simulations. Similarly, the four examinees with total test scores of 70.59% or 71.77%, representing less than 4% of the examinees, accounted for over 75% of the false master decisions across all the testing algorithms.

The fact that these data deviate significantly from the normal distribution rules out the use of statistical analysis approaches such as the repeated measures one-way analysis of variance (RM-ANOVA) that are based on the assumption that data are normally distributed. Fortunately, nonparametric alternatives are available that serve similar functions. A Friedman Test was conducted to examine differences in total error rates among the four algorithms, $\chi^2(3) = 11.22$, $p = .011$. The p value of the Friedman Test was slightly higher than the set value for individual hypothesis testing of $p = .01$ to indicate a significant result, and therefor a significant result was not found, and post-hoc testing was not conducted.

Table 19 presents the proportion of reduction in error (PRE) associated with each of the testing algorithms. PRE is just a linear transformation of the same data used to calculate

the overall error rate (i.e. correct decisions rate, false nonmaster rate, and false master rate), which was examined for significant differences across testing algorithms. Consequently, no statistical tests were conducted to determine whether or not differences between PRE values across algorithms were significant.

Table 19. *COM Test Simulations Proportion of Reduction in Error Descriptive Statistics and Test of Normality*

					Test of Normality		
Algorithm	*M*	*SD*	Skewness	Kurtosis	Shapiro-Wilk	df	*p*
SPRT	0.95	0.13	-3.15	10.72	.50	104	< .001
ARCH Pre-Cal	0.93	0.16	-3.29	12.13	.49	104	< .001
ARCH Post-Cal	0.91	0.29	-2.98	9.70	.53	104	< .001
EXSPRT	0.93	0.15	-3.26	11.43	.50	104	< .001

The higher the PRE value, the more accurate was a given testing algorithm with a PRE value of 1 representing a complete elimination of error. Examination of table 18 and 19 shows the expected inverse relationship between mean PRE values and mean overall error rates (due to how PRE is calculated).

False Nonmaster and False Master Error Rates

Tables 20 and 21 present the mean percentages and tests of normality for false nonmaster and false master rates.

Table 20. *COM Test Simulations False Nonmastery Rate Descriptive Statistics and Test of Normality*

					Test of Normality		
Algorithm	*M*	*SD*	Skewness	Kurtosis	Shapiro-Wilk	df	*p*
SPRT	1.69	5.15	4.32	21.23	.38	104	< .001
ARCH Pre-Cal	2.25	7.06	4.44	22.25	.37	104	< .001
ARCH Post-Cal	3.17	8.86	4.02	17.94	.41	104	< .001

EXSPRT	1.96	6.07	4.76	26.19	.37	104	< .001

Table 20 provides the mean false nonmaster rates across each of the four testing algorithms, along with standard deviation, skewness, kurtotois, and test of normality statistics. All the algorithms, with the exception of ARCH post-calibration, yielded mean false nonmaster error rates that were lower than the *a priori* rate of 2.5%. However, claims cannot be made regarding whether or not the 3.17% ARCH post-calibration false nonmaster rate is significantly higher than the 2.5% rate established *a priori*. Recall that results of chi-squared goodness-of-fit tests reported earlier demonstrated that 96% of the ARCH post-calibration tests did not yield error rates that deviated significantly from rates established *a priori*.

Given that, once again, skewness, kurtosis, and test of normality values indicates that the false nonmaster rate data are not normally distributed, the nonparametric Friedman Test was used to examine differences in false nonmaster rates among the four algorithms. The Friedman Test found a statistically significant difference in false nonmastery rates depending on which algorithm was used, $\chi^2(3) = 17.95$, $p < .001$. Post hoc analyses with six Wilcoxon signed-rank tests were conducted with a Bonferroni correction applied, resulting in a significance level set at $p < .0017$. There was a statistically significant reduction, $Z = -3.68$, $p < .001$, in the false nonmaster rate for SPRT ($M = 1.69$, $SD = 5.15$) versus ARCH post-calibration ($M = 3.17$, $SD = 8.86$), which is associated with an effect size of .36. The remaining comparisons did not reveal differences that were significant at the $p < .0017$ level. Significant reductions were not found between: (1) SPRT and ARCH pre-calibration, $Z = -1.95$, $p = .05$; (2) SPRT and EXSPRT, $Z = -0.484$, $p = .63$; (3) ARCH pre-calibration and

ARCH post-calibration, $Z = -2.48$, $p = .013$; (4) ARCH pre-calibration and EXSPRT, $Z = -0.825$, $p = .41$; and (5) ARCH post-calibration and EXSPRT, $Z = -2.91$, $p = .004$.

Table 21. *COM Test Simulations False Mastery Rate Descriptive Statistics and Test of Normality*

					Test of Normality		
Algorithm	*M*	*SD*	Skewness	Kurtosis	Shapiro-Wilk	df	*p*
SPRT	1.08	4.25	5.18	30.52	.49	104	< .001
ARCH Pre-Cal	1.10	4.46	4.88	25.00	.49	104	< .001
ARCH Post-Cal	1.40	5.45	4.44	19.13	.49	104	< .001
EXSPRT	1.33	5.02	4.78	23.61	.46	104	< .001

Table 19 presents the mean false master rate across each of the four testing algorithms along with standard deviation, skewness, kurtosis, and test of normality statistics. All algorithms had mean false master rates that were below the 2.5% false master rate established *a priori*. A Friedman Test was conducted to examine differences in false master rates among the four algorithms; however, no significant differences were found, $\chi^2(3) = 1.59$, $p = .662$, and therefor post-hoc testing was not conducted.

Summary

The results above address the research question – How accurate is ARCH in comparison to *a priori* error rates, SPRT, and EXSPRT? – in the context of Monte Carlo simulations with historical COM test data. Goodness-of-fit tests showed that ARCH pre-calibration did not have error rates that differed significantly from *a priori* error rates across all 50 sets of 104 tests. In the 50 ARCH post-calibration sets of 104 tests, 96% had error rates that did not differ significantly from error rates established *a priori*. Results of Friedman

Tests to examine if mean overall error, false nonmaster, and false master rates differed across SPRT, ARCH pre-calibration, ARCH post-calibration, and EXSPRT testing algorithms only found significant differences between the four testing algorithms with respect to false nonmaster rates. Subsequent post-hoc testing revealed only one significant difference – the mean ARCH post-calibration false nonmaster rate was significantly higher than the mean SPRT false nonmaster rate.

Results suggest that, overall, ARCH is an accurate testing approach whose error rates, in most cases, do not differ significantly from rates established *a priori* or from the error rates of SPRT or EXSPRT. However, in 4% of the test sets examined, ARCH post-calibration had error rates that did deviate significantly from *a priori* error rates. Furthermore, ARCH post-calibration was found to be significantly less accurate than SPRT with respect to mean false nonmaster rates.

5.2.2 RQ2 Results for ARCH Evaluation with New IU Plagiarism Test Examinees

Accuracy of ARCH was evaluated both before and after ARCH had been calibrated. A total of 1,202 unique examine tests were selected for analysis during the ARCH pre-calibration phase and 4,527 unique examinee tests were selected for analysis during the ARCH post-calibration phase.

Since EXSPRT served as the measure of the true classification of each examinee (master or nonmaster) in this study, the accuracy of EXSPRT is not included in the analysis below. Unlike the COM test, where all 85 questions were answered by each examinee, there was no known *total* test score for each examinee who took the IU Plagiarism Test which could then be compared with a cut-score to determine his or her mastery status. Thus, EXSPRT was chosen as the standard for comparison, because it had demonstrated accuracy

in predicting total test decisions in past studies (Frick, 1992; Welch & Frick, 1993) with prediction errors within theoretically expected ranges.

Also, examinees taking the IU Plagiarism Test were either in the ARCH pre-calibration or the ARCH post-calibration group. Consequently, ARCH pre-calibration and ARCH post-calibration are analyzed separately (since these are independent groups). The following agreement tables provide results for ARCH pre and post-calibration, SPRT, and EXSPRT. Second, results of goodness-of-fit tests are provided. Finally, results of Friedman Tests and subsequent post-hoc testing are presented that examined error rates to determine if differences observed between the testing algorithms were significant.

Agreement Tables

I constructed tables to show how ARCH pre-calibration and ARCH post-calibration agreed or disagreed with SPRT and EXSPRT.

Table 22. *ARCH Pre-Calibration Decision Agreement with SPRT (Percent Agreement in Parentheses)*

		ARCH Pre-Calibration Decision					
		Nonmaster		Master		No Decision	
SPRT Decision	*n*	*n* = 729		*n* = 325		*n* = 148	
Nonmaster	729	729	(100)	0	(0)	0	(0)
Master	325	0	(0)	325	(100)	0	(0)
No Decision	148	0	(0)	0	(0)	148	(100)

ARCH pre-calibration agreed perfectly with SPRT, as expected. This is not surprising given that ARCH pre-calibration mimics SPRT and, unlike the first study, the testing algorithms in the second study used items in the order actually administered in real-time to examinees. Had these two sets of decisions not agreed perfectly here, this would have been an indication of a software error during the Web-based Plagiarism Test administrations.

Table 23. *ARCH Pre-Calibration Decision Agreement with EXSPRT (Percent Agreement in Parentheses)*

		ARCH Pre-Calibration Decision					
		Nonmaster		Master		No Decision	
EXSPRT Decision	*n*	*n* = 729		*n* = 325		*n* = 148	
Nonmaster	653	630	(86.42)	6	(1.85)	17	(11.49)
Master	549	99	(13.58)	319	(98.15)	131	(88.51)

ARCH pre-calibration did not agree perfectly with EXSPRT. While ARCH pre-calibration and EXSPRT agreed 98.15% of the time with respect to master decisions, they only agreed 86.42% of the time with respect to nonmaster decisions. The relatively low rate of agreement between ARCH pre-calibration and EXSPRT with respect to nonmaster decisions is reflected in the high false nonmaster rate of ARCH pre-calibration, which will be detailed later in this chapter. Interestingly, in the 148 cases when ARCH pre-calibration was not able to make a decision, EXSPRT classified most of them (88.51%) as masters.

Table 24. *ARCH Post-Calibration Decision Agreement with SPRT (Percent Agreement in Parentheses)*

		ARCH Post-Calibration Decision					
		Nonmaster		Master		No Decision	
SPRT Decision	*n*	*n* = 2615		*n* = 1704		*n* = 208	
Nonmaster	2863	2514	(96.14)	257	(15.08)	92	(44.23)
Master	1132	31	(1.19)	1083	(63.56)	18	(8.65)
No Decision	532	70	(2.68)	364	(21.36)	98	(47.12)

Unlike ARCH pre-calibration, ARCH post-calibration did not agree perfectly with SPRT. ARCH post-calibration agreed with SPRT in 96.14% of nonmaster decisions but only agreed with SPRT in 63.56% of master decisions. When ARCH was not able to make a

decision, SPRT was nearly equally likely to make a nonmaster decision (44.23%) or not make a decision (47.12%).

Table 25. *ARCH Post-Calibration Decision Agreement with EXSPRT (Percent Agreement in Parentheses)*

		ARCH Post-Calibration Decision					
		Nonmaster		Master		No Decision	
EXSPRT Decision	*n*	*n* = 2615		*n* = 1704		*n* = 208	
Nonmaster	2613	2432	(93.00)	92	(5.40)	89	(42.79)
Master	1914	183	(7.00)	1612	(94.60)	119	(57.21)

ARCH post-calibration had high levels of agreement with EXSPRT. ARCH post-calibration agreed with EXSPRT in 93% of nonmaster decisions and 94.60% of master decisions. When ARCH post-calibration could not make a decision EXSPRT was slightly more like to classify that individual as a master. Overall, this finding is consistent with theoretical expectations, based on the thresholds established for when the ARCH post-calibration method starts being used with test examinees (determined in the Monte Carlo simulations conducted earlier).

Chi-Squared Goodness-of-fit

Two chi-squared goodness-of-fit tests were conducted to examine if ARCH pre-calibration and ARCH post-calibration made classification decisions within *a priori* error rates. Recall that EXSPRT calibrated with 50 nonmasters and 50 masters was used to represent the true score, since the participants in the study only ever answered a small subset of the items available in the item pool. The *a priori* false nonmaster and false master rates were both set to 5% for the new IU Plagiarism Test.

The first goodness-of-fit test found that the observed results of ARCH pre-calibration testing deviated significantly, $\chi^2=(2, N = 1054) = 82.06$, $p < .001$, from error rates set *a*

priori. EXSPRT was used in this study to represent the true examinee classification. As shown in table 26, ARCH pre-calibration agreed with SPRT 100% of the time. Consequently, the significant lack of fit between ARCH pre-calibration and the *a priori* error rate is due to a lack of fit between SPRT and EXSPRT decisions.

The second goodness-of-fit test found that the observed results of ARCH pre-calibration testing also deviated significantly, χ^2=(2, N = 1054) = 82.51, $p < .001$, from error rates set *a priori*. In this case, ARCH post-calibration deviated from the *a priori* error rates because it made significantly *fewer* errors than expected *a priori*.

Overall Error Rate and Proportion of Reduction in Error

Tables 26 and 27 provide the overall error rate and PRE for ARCH pre-calibration and post-calibration respectively. The overall error rate and proportion of reduction in error (PRE) is the same for both SPRT and ARCH pre-calibration. Both made 1,054 decisions, of which 949 (90.04%) were correct and 105 (9.96%) were errors. The 9.96% overall error rate is below the total *a priori* error rate of 10%. The PRE for both was .8008. As SPRT and ARCH pre-calibration made identical decisions and had identical error rates, no tests were conducted to determine if overall error rates differed. Again, as explained above, these two methods should perform exactly the same, assuming that the Web-based test administration software is working correctly.

Table 26. *IU Plagiarism Test Pre-Calibration Overall Error Rate and Proportion Reduction in Error*

Algorithm	*Errors*	*Decisions*	% Correct	% Errors	PRE
SPRT	105	1,054	90.04	9.96	.8008
ARCH Pre-Cal	105	1,054	90.04	9.96	.8008

Table 27. *IU Plagiarism Test Post-Calibration Overall Error Rate and Proportion Reduction in Error*

Algorithm	*Errors*	*Decisions*	% Correct	% Errors	PRE
SPRT	366	3,995	90.84	9.16	.8168
ARCH Post-Cal	275	4,319	93.63	6.37	.8726

SPRT made 3,995 decisions, of which 3,629 (90.84%) were correct and 366 (9.16%) were errors. The PRE for SPRT was .8168. ARCH post-calibration made 4,319 decisions, of which 4,044 (93.63%) were correct and 275 (6.37%) were errors. The PRE for ARCH post-calibration was .8726. A Wilcoxon Signed Rank test found that there was a significant reduction in overall error, $Z = -6.95$, $p < .001$, between ARCH post-calibration and SPRT with an effect size of .11.

False Nonmaster and False Master Rates

Tables 28 and 29 provide the false nonmaster error rate for ARCH pre-calibration and post-calibration respectively.

Table 28. *IU Plagiarism Test Pre-Calibration False Nonmaster Error Rate*

Algorithm	*False Nonmaster Errors*	*Decisions*	% False Nonmaster
SPRT	99	1,054	9.39
ARCH Pre-Cal	99	1,054	9.39

Both SPRT and ARCH pre-calibration made 1,054 decisions, of which 99 (9.39%) were false nonmaster errors, which is above the false nonmaster *a priori* error rate of 5%. As SPRT and ARCH pre-calibration made identical decisions and, consequently, had identical false nonmaster error rates, no tests were conducted to determine if overall error rates differed.

Table 29. *IU Plagiarism Test Post-Calibration False Nonmaster Error Rate*

Algorithm	*False Nonmaster Errors*	*Decisions*	% False Nonmaster
SPRT	348	3,995	8.71
ARCH Post-Cal	183	4,319	4.24

Of the 3,995 decisions made by SPRT during the post-calibration phase, 348 (8.71%) were false nonmaster errors, which is above the 5% false nonmaster rate established *a priori*. ARCH post-calibration made 4,319 decisions, of which 183 (4.24%) were false nonmaster errors, which is below the 5% false nonmaster rate established *a priori*. A Wilcoxon Signed Rank test found that there was a significant reduction in false nonmaster errors, $Z = -11.54$, $p < .001$, between ARCH post-calibration and SPRT with an effect size of .19.

Tables 30 and 31 provide the false master error rate for ARCH pre-calibration and post-calibration respectively.

Table 30. *IU Plagiarism Test Pre-Calibration False Master Error Rate*

Algorithm	*False Master Errors*	*Decisions*	% False Master
SPRT	6	1,054	0.57
ARCH Pre-Cal	6	1,054	0.57

Both SPRT and ARCH pre-calibration made 1,054 decisions, of which 6 (0.57%) were false master errors, which is well below the false master *a priori* error rate of 5%. As SPRT and ARCH pre-calibration made identical decisions and, consequently, had identical false master error rates, no tests were conducted to determine if overall error rates differed.

Table 31. *IU Plagiarism Test Post-Calibration False Master Error Rate*

Algorithm	*False Master Errors*	*Decisions*	% False Master
SPRT	18	3,995	0.45
ARCH Post-Cal	92	4,319	2.13

Of the 3,995 decisions made by SPRT during the post-calibration phase, 18 (0.45%) were false master errors, which is well below the 5% false master rate established *a priori*. ARCH post-calibration made 4,319 decisions, of which 92 (2.13%) were false master errors, which is well below the 5% false master rate established *a priori*. A Wilcoxon Signed Rank test found that there was a significant reduction in false master errors, $Z = -6.65$, $p < .001$, between ARCH post-calibration and SPRT with an effect size of .11. Thus, SPRT made fewer false master decisions than expected, but apparently at the expense of making many more false nonmaster decisions (see tables 23 and 29). This means that a SPRT decision for mastery was correct nearly all of time, but a SPRT decision of nonmastery was correct only 86 percent of the time.

Summary

The results above address the research question – How accurate is ARCH in comparison to *a priori* error rates, SPRT, and EXSPRT? – in the context of new examinees who took a new version of the IU Plagiarism Test. Agreement tables showed that ARCH pre-calibration agreed 100% of the time with SPRT but that ARCH pre-calibration/SPRT made nonmaster decisions that only agreed with EXSPRT in 86.42% of cases. ARCH post-calibration had high levels of agreement with SPRT for nonmaster decisions but only agreed with SPRT in 63.56% of master decisions and 47.12% of no-decisions.

ARCH post-calibration had high levels of agreement with EXSPRT for both nonmaster and master decisions. Goodness-of-fit testing showed that ARCH pre-calibration and ARCH post-calibration had error rates that differed significantly from *a priori* error rates. However, ARCH pre-calibration's lack of fit with *a priori* error rates can be explained

by the fact that ARCH pre-calibration was identical to SPRT, and SPRT decisions differed from EXSPRT, which was used to represent the true examinee state. ARCH post-calibration's lack of fit with *a priori* error rates is due to ARCH post-calibration having error rates lower than those expected. In the 50 ARCH post-calibration sets of 104 tests, 96% had error rates that did not differ significantly from error rates established *a priori*. SPRT and ARCH pre-calibration made identical false nonmaster errors with an error rate that was above the 5% false nonmaster rate established *a priori*. A Wilcoxon Signed Rank test found ARCH post-calibration made significantly fewer overall errors and false nonmaster errors than SPRT. While both SPRT and ARCH post-calibration had false master error rates below the 5% level established *a priori*, Wilcoxon Signed Rank testing found ARCH post-calibration made significantly more false master errors than did SPRT.

Results suggest that SPRT/ARCH pre-calibration decision differed significantly from EXSPRT, which was used to represent true examinee classification. ARCH post-calibration was found to be an accurate testing approach that differed significantly from rates established *a priori* because of lower than expected error rates. ARCH post-calibration did make significantly more false mastery decisions than did SPRT but made significantly fewer false nonmaster errors.

5.3 ARCH Efficiency (RQ3) Results

The research question – How efficient is ARCH in comparison to traditionally calibrated SPRT and EXSPRT? – was answered in both the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data, and the second study, ARCH evaluation with new IU plagiarism test examinees.

5.3.1 RQ3 Results for Monte Carlo ARCH Evaluation with Historical COM Test Data

The assumption of normality of the test lengths and no-decision rates was evaluated for SPRT, ARCH pre-calibration, ARCH post-calibration, and EXSPRT via examination of the skewness and kurtosis values of the variables and through the Shapiro-Wilk test of normality, since the number of observations is less than 200. Results presented in Tables 32 and 33 show that normality cannot be assumed across all the variables examined.

The test length data and no-decision rates associated with each of the algorithms deviated from a normal distribution curve. Each algorithm's test length data and no-decision data were substantially positively skewed, which indicates a far from symmetrical distribution and a higher frequency of shorter test lengths. For example, Kurtosis values of SPRT and ARCH post-calibration provide evidence that the underlying distributions are more peaked than would be expected if the data conformed to the normal distribution. Most convincingly, the Shapiro-Wilk test of normality conducted on each of the test length and no-decision rate data sets yielded significant results ($p <= .001$) in all cases, which indicates that the data differ significantly from the normal distribution, so nonparametric tests needed to be employed.

Table 32. *COM Test Length Descriptive Statistics and Tests of Normality*

					Test of Normality		
Algorithm	*M*	*SD*	Skewness	Kurtosis	Shapiro-Wilk	df	*p*
SPRT	22.45	10.79	1.18	.80	.89	104	< .001
ARCH Pre-Cal	21.98	9.97	.99	.12	.93	104	<.001
ARCH Post-Cal	13.90	4.93	.92	1.01	.90	104	< .001
EXSPRT	13.42	5.29	.80	-.14	.95	104	.001

The nonparametric Friedman Test was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.01$ (.05 divided by 5). The Friedman Test examined differences in test lengths among the four algorithms. The Friedman Test found a statistically significant difference in test length depending on which algorithm was used, $\chi^2(3) = 254.57$, $p < .001$. Post hoc analysis with six Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.0017$.

There were no significant differences between the ARCH pre-calibration and SPRT algorithm test lengths, $Z = -1.65$, $p = .098$. However, statistically significant differences were found among the remaining five algorithm comparisons. There was a statistically significant reduction, $Z = -8.85$, $p < 0.001$, in test length for ARCH post-calibration ($M = 13.90$, $SD = 4.93$) versus SPRT ($M = 22.45$, $SD = 10.79$). Mean test lengths for ARCH post-calibration ($M = 13.90$, $SD = 4.93$) were also significantly shorter, $Z = -8.85$, $p < 0.001$, than ARCH pre-calibration ($M = 21.98$, $SD = 9.97$). EXSPRT mean test lengths ($M = 13.42$, $SD = 5.29$) were found to be significantly shorter, $Z = -8.85$, $p < 0.001$, than SPRT ($M = 22.45$, $SD = 10.79$), significantly shorter, $Z = -8.85$, $p < 0.001$, than ARCH pre-calibration ($M = 21.98$, $SD = 9.97$), and, surprisingly, significantly shorter, $Z = -3.78$, $p < 0.001$, than ARCH post-calibration ($M = 13.90$, $SD = 4.93$). The EXSPRT versus ARCH post-calibration mean test length comparison was associated with an effect size of .37 with the remaining significant differences associated with the large effect size of .87.

The repeated occurrence of $Z = -8.85$ in the results above are due to the fact that for all 104 examinees both EXSPRT and ARCH post-calibration had shorter mean test lengths than both SPRT and ARCH pre-calibration. Mean EXSPRT test lengths were shorter than

ARCH post-calibration for 70 examinees and longer in the remaining 34 cases, which resulted in the significant Wilcoxon signed-rank test, despite the small overall difference in the mean of means of just 0.48.

Table 33. *COM Test No-Decision Rates and Test of Normality*

					Test of Normality		
Algorithm	*M*	*SD*	Skewness	Kurtosis	Shapiro-Wilk	df	*p*
SPRT	2.00	5.52	2.97	8.08	.48	104	< .001
ARCH Pre-Cal	1.81	4.98	3.33	11.21	.45	104	< .001
ARCH Post-Cal	0.17	0.79	5.50	33.53	.53	104	< .001
EXSPRT	0.04	0.39	10.20	104.00	.53	104	< .001

The Friedman Test examined differences in mean no-decision rates among the four algorithms. The Friedman Test found a statistically significant difference in false nonmastery rates depending on which algorithm was used, $\chi^2(3) = 42.45$, $p < .001$. Post hoc analysis via six Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, using the previously established significance level of $p < .0017$. Significant reductions in no-decision rates were not found between: (1) SPRT and ARCH pre-calibration, $Z = -0.64$, $p = 0.52$, and (2) ARCH post-calibration and EXSPRT, $Z = -1.47$, $p = 0.14$.

Statistically significant reductions in no-decision rates were found for the rest of the comparisons. There was a statistically significant reduction, $Z = -3.63$, $p < 0.001$, in the no-decision rate for SPRT ($M = 2.00$, $SD = 5.52$) versus ARCH post-calibration ($M = 0.17$, $SD = 0.79$), which is associated with an effect size of .36. SPRT ($M = 2.00$, $SD = 5.52$) had a significantly higher, $Z = -3.63$, $p < 0.001$, no-decision rate compared to EXSPRT ($M = 0.04$, $SD = 0.39$), which is associated with an effect size of .36. There was a significant reduction,

Z = -3.82, $p < 0.001$, in no-decision rate from ARCH pre-calibration (M = 1.81, SD = 4.98) versus ARCH post-calibration (M = 0.17, SD = 0.79), which was associated with an effect size of .37. Finally, ARCH pre-calibration (M = 1.81, SD = 4.98) had a significantly higher, Z = -3.93, $p < 0.001$, no-decision rate compared to EXSPRT (M = 0.04, SD = 0.39), which is associated with an effect size of .39.

Summary

Results presented above answer the research question – How efficient is ARCH in comparison to SPRT and EXSPRT – in the context of Monte Carlo simulations with COM test data. Overall findings suggest that, as expected, ARCH pre-calibration did not differ significantly from SPRT but that ARCH post-calibration provides significant reductions in test length and no-decision rates when compared to SPRT. ARCH post-calibration did not differ from EXSPRT in terms of no-decision rates, but the small 0.48 increase in mean test lengths associated with ARCH-post calibration when compared to EXSPRT was found to be significant.

5.3.2 RQ3 Results for ARCH Evaluation with New IU Plagiarism Test Examinees

The following analyzes the efficiency of ARCH in the context of the new IU Plagiarism Test. Mean test lengths and no-decision rates serve as measures of test efficiency. In this study, if none of the algorithms could make a mastery or nonmastery decision after 20 items were administered, using the *a priori* classification error rates, then the test ended in a no-decision classification for this examinee. While this fact was recorded in the database, an examinee was informed at the end of 20 questions that no clear mastery decision could be confidently reached. From a practical perspective, she or he had not passed the test—since a no-decision or nonmastery decision was still not a mastery decision.

Examinees taking the IU Plagiarism Test either took the test using the ARCH pre-calibration method (equivalent to SPRT) or the ARCH post-calibration method. Consequently, ARCH pre-calibration and ARCH post-calibration methods are analyzed separately.

Table 34. *IU Plagiarism Test Length Descriptive Statistics and Tests of Normality for ARCH Pre-Calibration*

					Test of Normality		
Algorithm	*M*	*SD*	Skewness	Kurtosis	Shapiro-Wilk	df	*p*
SPRT	9.96	5.83	0.58	-1.04	.874	1202	< .001
ARCH Pre-Cal	9.96	5.83	0.58	-1.04	.874	1202	< .001
EXSPRT	7.22	4.39	1.06	0.46	.898	1202	<.001

Since SPRT and ARCH pre-calibration have identical test length descriptive statistics, only a pairwise test was conducted to determine if differences were significant between mean test lengths of SPRT/ARCH pre-calibration and EXSPRT. A Wilcoxon signed-rank test found a statistically significant reduction, $Z = -19.46$, $p < 0.001$, in the test length for SPRT/ARCH pre-calibration ($M = 9.96$, $SD = 5.83$) versus EXSPRT ($M = 7.22$, $SD = 4.39$), which is associated with an effect size of .56.

Table 35. *IU Plagiarism Test Length Descriptive Statistics and Tests of Normality for ARCH Post-Calibration*

					Test of Normality		
Algorithm	*M*	*SD*	Skewness	Kurtosis	Shapiro-Wilk	df	*p*
SPRT	10.05	5.81	0.53	-1.10	.88	4527	< .001
ARCH Post-Cal	8.15	5.03	0.93	-0.05	.89	4527	< .001
EXSPRT	7.10	4.31	0.99	0.32	.91	4527	<.001

The Friedman Test examined differences in test lengths among the three algorithms. The Friedman Test found a statistically significant difference in test lengths depending on which algorithm was used, $\chi^2(2) = 2025.91$, $p < .001$. Post hoc analysis via three Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < .0017$. A Wilcoxon signed-rank test found a statistically significant reduction, $Z = -39.36$, $p < 0.001$, in the test length for SPRT ($M = 10.05$, $SD = 5.81$) versus EXSPRT ($M = 7.10$, $SD = 4.31$), which is associated with an effect size of .58. A Wilcoxon signed-rank test found a statistically significant reduction, $Z = -26.37$, $p < 0.001$, in the test length for SPRT ($M = 10.05$, $SD = 5.81$) versus ARCH post-calibration ($M = 8.15$, $SD = 5.03$), which is associated with an effect size of .39. A Wilcoxon signed-rank test found a statistically significant reduction, $Z = -18.41$, $p < 0.001$, in the test length for ARCH post-calibration ($M = 8.15$, $SD = 5.03$) versus EXSPRT ($M = 7.10$, $SD = 4.31$), which is associated with an effect size of .27.

Table 36. *IU Plagiarism Test Pre-Calibration No-Decision Rate*

Algorithm	*No Decisions*	*Tests*	% No Decision
SPRT	148	1202	12.31
ARCH Pre-Cal	148	1202	12.31

Both SPRT and ARCH pre-calibration could not make a classification decision in 148 of the 1,202 tests, for a no-decision rate of 12.31%. As SPRT and ARCH pre-calibration made identical decisions and, consequently, had identical no-decision rates, no tests were conducted to determine if overall error rates differed.

Table 37. *IU Plagiarism Test Post-Calibration False Nonmaster Error Rate*

Algorithm	*No Decisions*	*Tests*	% No Decision
SPRT	532	4,527	11.75
ARCH Post-Cal	208	4,527	4.59

SPRT could not make a classification decision in 532 of the 4,527 tests for a no-decision rate of 11.75%. ARCH pre-calibration could not make a classification decision in 208 of the 4,527 tests for a no-decision rate of 4.59%. A Wilcoxon Signed Rank test found that there was a significant reduction in the no-decision rate, $Z = -13.89$, $p < .001$, between ARCH post-calibration and SPRT with an effect size of .27.

Summary

Results presented above answer the research question – How efficient is ARCH in comparison to SPRT and EXSPRT – in the context of testing with new examinees associated with a new version of the IU Plagiarism Test. Overall findings suggest that, as expected, ARCH pre-calibration behaved exactly as SPRT with identical test lengths, but that ARCH post-calibration provided significant reductions in test length and no-decision rates when compared to SPRT. ARCH post-calibration was found to have significantly longer test lengths when compared to EXSPRT, however, this difference was approximately one additional item on average.

CHAPTER VI. DISCUSSION

The three sections of this chapter present and discuss the major findings and implications associated with each of the three research questions. RQ1 was answered exclusively in the first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data. RQ2 and RQ3 were answered in both the first and the second study. ARCH was evaluated in the latter study with new IU plagiarism test examinees. The chapter concludes with discussion of the finding that SPRT had a higher false nonmaster error rate than expected, followed by a broader discussion of viability of the ARCH approach in real-world contexts and the practical implications of this research.

6.1 ARCH Calibration Sufficiency (RQ1)

RQ1 – When are items sufficiently calibrated? – was addressed in the first study. An item was found to be sufficiently calibrated when two item calibration criteria thresholds and one item-bank level criterion threshold were met. The two item calibration criteria were: (1) Beta standard deviation values associated with both the probability of a correct answer to the item from a master, $P(C_i|M)$, and the probability of a correct answer to the item from a nonmaster, $P(C_i|NM)$, had to reach or be less than .078 to indicate sufficient precision of the beta mean estimate; (2) A *Beta Difference Index* value, a measure of the difference between $P(C_i|M)$ and $P(C_i|N)$, had to reach or be greater than 15 to indicate a sufficient difference. Finally, the item-bank level criterion required that at least 83% of items in the pool had to meet the two item-level criteria above before ARCH could begin (adaptive testing using item-level calibration data).

This set of calibration criteria thresholds provided an answer to RQ1, but the process of establishing the answer to RQ1 also yielded noteworthy findings. The novel use of beta

SD in the context of item calibration and the invention of the *Beta Difference Index* statistic also represent key findings of my research. The implications of the establishment of a set of calibration criteria thresholds and new item calibration and quality statistics are, in turn, discussed below.

6.1.1 Viable Set Of Calibration Criteria Thresholds

The set of calibration criteria thresholds established through thousands of Monte Carlo simulations not only provide values required for the heuristics that make up ARCH, but also have the potential to be useful for determining test accuracy and efficiency in other assessment contexts. The brute force approach applied in the first study for determining the set of calibration criteria threshold values that would lead to efficient and accurate VL-CCT in Monte Carlo simulations with COM test data was necessary because of the limited research into item calibration for Classical Test Theory-based VL-CCT approaches. Frick's (1992) study was the lone study to provide any guidance on the level of calibration data necessary to facilitate efficient and accurate EXSPRT-based testing.

Going forward, the set of calibration criteria thresholds established in my research will provide a starting point for subsequent investigations in other contexts. The set of calibration criteria thresholds established in this study might be found to be overly strict or too lenient using different testing data. Furthermore, it may be that fewer or more calibration criteria are necessary to reliably predict when items are sufficiently calibrated for use with VL-CCT approaches based on Classical Test Theory once more testing contexts are examined.

The fact that the set of calibration criteria thresholds established through Monte Carlo simulations based on COM test data worked effectively in the very different context of the

IU Plagiarism Test does provide evidence of the robustness of these thresholds. The COM test was comprised of 85 items and proved to be a much easier test for examinees than was the IU Plagiarism Test with a much larger pool of considerably more difficult items. For example, nonmasters on the COM test were, on average, 33% more likely to answer an item correctly compared to the IU Plagiarism Test. In addition, simulations using the COM test did not have any restriction on the maximum test length, whereas the IU Plagiarism Test limited test lengths to 20 items. A further difference between the two testing contexts was that the *a priori* error rates with the IU Plagiarism Test were twice as large as those used in COM test Monte Carlo simulations. Despite these differences in testing contexts, the set of calibration criteria thresholds led to efficient and accurate testing with ARCH post-calibration in both cases.

While developed specifically for ARCH, the set of calibration criteria thresholds can be applied in other adaptive and non-adaptive testing contexts to evaluate calibration sufficiency or item quality. Traditionally, calibration associated with EXSPRT involves the collection of item calibration data during a separate calibration phase before adaptive testing methods are deployed (Frick, 1992; Welch & Frick, 1993; Welch, 1997). The set of calibration criteria thresholds established in this study could be applied in a separate item calibration phase to indicate when items have been sufficiently calibrated, rather than limiting calibration to a fixed number of examinees.

A more far-reaching potential use of the set of calibration criteria thresholds involves adding to or replacing existing statistics and thresholds used in item quality analysis. For example, the index of discrimination and associated interpretation guidelines from Ebel (1972) are a widely taught item quality statistic in assessment and evaluation textbooks (e.g.,

Reynolds et al., 2010) used to examine item quality but do not factor in the number of observations. In other words, a given index of discrimination value may be based on very few observations and, consequently, may not be sufficiently reliable to base decisions on (e.g., whether or not to eliminate or revise the item).

In comparison, the *Beta Difference Index* and associated threshold value of 15 does factor in the number of observations. Therefore, use of the *Beta Difference Index* would provide an empirically justified basis for decisions about elimination or revision of items. Selection of masters and nonmasters on tests that are not classification-focused could apply Kelley's (1939) method for selecting upper and lower ability groups where the highest 27% of scores make up the upper ability group and the lowest 27% of scores make up the lower ability group. While more complicated to compute than the index of discrimination, the *Beta Difference Index* can easily be calculated either through the use of Excel or the development of a relatively simple online equation tool.

6.1.2 New Item Calibration and Quality Statistics

Statistics initially proposed to indicate calibration sufficiency – a measure of the precision of item calibration estimates and a measure of the differences between *P(C|M)* and *P(C|NM)* – were abandoned in favor of two less computationally complex and more understandable alternatives. Beta *SD* is an existing statistic that has never been applied to the problem of item calibration sufficiency. The *Beta Difference Index* is a new statistic I derived based on existing equations for beta distributions and Welch's *t*-test equation: to address the need to efficiently evaluate the degree of difference between two beta distributions. Both the novel application of beta *SD* and the invention of the *Beta Difference Index* are major findings of my research.

Use of the beta *SD* equation to measure the precision of a probability estimate of a correct answer from a specific classification group has, to my knowledge, not been done prior to my research. Originally, I had proposed using the 95% Highest Density Region Width (HDRW) associated with a given beta distribution to measure the precision of a probability estimate. However, the numerical integration method required for determining the HDRW was found to be impractical for deployment in massively open online testing contexts due to computational demand required to dynamically calculate the value. The beta *SD* proved to be an excellent replacement for the HDRW since it is straightforward to compute and correlates very strongly with HDRW values.

Similarly, the new *Beta Difference Index* proved to be an excellent alternative to the originally proposed shared area under beta curve methods to measure difference between the beta distributions associated with the probability of a correct answer from a nonmaster and that of a master. As a new type of statistical test for examining the differences between two beta distributions, it is not clear if specific *Beta Difference Index* values should be associated with particular *p* values, under what conditions the test operates reliably, or what assumptions should be associated with the test. While the *Beta Difference Index* formula is inspired by *t*-test formulas, it is not clear if the proposed *Beta Difference Index* conforms to a *t*-distribution. Critical examination by experts in statistics is needed to address these questions.

6.2 ARCH Accuracy (RQ2)

RQ2 was addressed in both study 1 and study 2. In most cases, ARCH made classification decisions that did not differ significantly from *a priori* error rates, SPRT error rates, or EXSPRT error rates. However, in some circumstances ARCH did have significantly

higher error rates than expected *a priori* and did differ significantly from SPRT and EXSPRT error rates.

6.2.1 Monte Carlo ARCH Evaluation with Historical COM Test Data

Overall, in analyses of Monte Carlo simulations using historical COM test data, ARCH was found to be an accurate testing approach. With two exceptions, ARCH error rates did not differ significantly from rates established *a priori* or from the error rates of SPRT or EXSPRT methods. First, a small percentage (4%) of the test sets for ARCH post-calibration had error rates that did deviate significantly from *a priori* error rates. Second, the ARCH post-calibration method resulted in a significantly higher false nonmaster rate than did the SPRT.

6.2.2 ARCH Evaluation with New IU Plagiarism Test Examinees

Results of the second study lead to very different conclusions for ARCH pre-calibration and ARCH post-calibration methods. As expected, ARCH pre-calibration decisions exactly matched SPRT decisions, but differed significantly from those of the EXSPRT method, which was used to represent true examinee classification. The ARCH post-calibration method was found to be an accurate testing approach that differed significantly from rates established *a priori* only because error rates were *lower than expected*. The ARCH post-calibration method error rates also highly agreed with those of EXSPRT. On the other hand, the ARCH post-calibration method made significantly more false mastery decisions than did SPRT, but the ARCH post-calibration false master rate was nonetheless *below* the *a priori* error rate. ARCH post-calibration made significantly *fewer* false nonmaster errors than did the SPRT method.

6.2.3 Discussion of ARCH Accuracy (RQ2)

The overall finding that, in most cases, ARCH is an accurate VL-CCT approach is consistent with results investigating other instances of the item-level application of Classical Test Theory such as Frick's EXSPRT (1992) and Rudner's Measurement Decision Theory (2002). The specific situations where ARCH did not achieve expected accuracy rates are discussed in the context of related literature below.

Goodness-of-fit findings for ARCH from the two studies are very different in terms of when and why pre- and post-calibration ARCH deviated significantly from *a priori* rates. ARCH pre-calibration had a good fit with expected error rates in the first study, but deviated significantly in the second study, which yielded higher-than-expected error rates. ARCH post-calibration had a few deviations from expected error rates in the first study due to *higher* than expected error rates, but deviated from expected error rates in the second due to the opposite reason – *lower*-than-expected error rates.

One explanation for the finding that 4% of test sets for ARCH post-calibration had error rates that deviated significantly from *a priori* error rates in the first study is that discovering at least one significant result due to a type 1 error was nearly guaranteed. With 50 chi-squared goodness-of-fit tests for ARCH pre-calibration test sets and a p value set to .05, the overall chances of committing at least one type I error was higher than 4%. As discussed earlier, the rationale for keeping the .05 p values for the chi-squared goodness-of-fit tests was to err on the side of caution and increase the likelihood that a lack of fit would be detected.

The reason for ARCH pre-calibration's lack of fit with expected error rates due to higher than expected error rates in the second study is the unexpectedly high error rate of

SPRT – the testing algorithm used during ARCH pre-calibration. The SPRT has been criticized historically for use as a computer adaptive testing (CAT) method (e.g., see Reckase, 1983; Ferguson, 1969; Weiss & Kingsbury, 1983) because it does *not* take into account differences in item difficulties or their discriminating power.

For example, if an examinee happens to get randomly selected questions early in the test that are easy and not very discriminating between masters and nonmasters, and if those questions were answered *correctly*, she or he would be classified as a master by the SPRT. Alternatively, if that *same examinee* happens to get questions that are more difficult and highly discriminating between masters and nonmasters, and answered those questions mostly *incorrectly*, she or he would be classified as a nonmaster by the SPRT method. This is the fundamental reason why Plew (1989) and Frick (1992) created the EXSPRT method—in order to take into account item difficulty levels, as well as their ability to discriminate masters from nonmasters—when computing probability ratios.

Given this limitation of the SPRT for use in CATs, Frick (1989) recommended that the SPRT be used conservatively—by not making the zone of indifference too wide, and by choosing very low *a prio*ri error rates—in order to keep tests from being too short. This is because shorter tests are more likely to result in classification errors when item difficulty and discrimination are not accounted for when computing probability ratios.

In the second study of the IU Plagiarism test, practical concerns drove the decisions for choosing a somewhat wider zone of indifference (.35) than that recommend by Frick (.25) and higher error rates for false mastery and nonmastery decisions (.05 instead of .025). The parameters chosen for SPRT in the second study here were a compromise, in order to keep

tests from being too long, with the prior knowledge that the passing rate for the Plagiarism Test was relatively low—meaning that it took most examinees a number of attempts before a test was passed. The reasoning was that false nonmastery decision errors were less important than false mastery decision errors. That is, it was better to provide a certificate for passing the test, when an examinee is a true master of recognizing plagiarism, compared with erroneously giving a certificate to one who was a actually a nonmaster.

Item response theory (IRT) was invented by Lord and Novick (1968) as a way of accounting for item difficulty, discrimination, and chances of guessing (referred to as the lower asymptote). While IRT has been demonstrated to work reasonably well with standardized tests when estimating an examinee's ability level, one major issue has been the large number of examinees (thousands) which are necessary to estimate item parameters *prior* to actually implementing an IRT-based CAT. This requirement is not practical for most instructional contexts, and why computerized classification tests (CCTs) have been subsequently considered as a more practical alternative to CAT.

The higher-than-expected false nonmaster error rate of SPRT in the first study with empirically established parameter estimates and in the second study with the IU Plagiarism Test is discussed in detail in a separate section later in this chapter. However, it is worth noting that similar findings are not reported in most of the previous studies examining the application of SPRT in educational testing contexts (Frick, 1989; Frick, 1992; Welch & Frick 1993), with studies by Plew (1989) and Frick (1990) being notable exceptions. Plew's dissertation study found that a variety of adaptive testing methods, including SPRT, did not perform within *a priori* error rates. Plew suggested that clustering of scores around the cut-off point was likely to blame for lower-than-expected SPRT accuracy, as did Frick (1990),

who found that SPRT did not perform within *a priori* error rates when many examinees straddled the boundary between nonmastery and mastery.

Related to the goodness-of-fit findings just discussed is the fact that in the first study ARCH post-calibration had a mean false nonmaster rate of 3.17% which is 0.67% higher than the 2.5% *a priori* rate and substancially higher than SPRT's mean false nonmaster rate. We cannot just dismiss the 4% of ARCH post-calibration sets of tests that deviated from *a priori* error rates in the first study as being solely due to type I error, as there is evidence that suggests that ARCH post-calibration was prone to making more false nonmaster errors than expected *a priori*. A central idea of Classical Test Theory, the basis for all the VL-CCT algorithms, is that an observed test score is made up of the true score and error (Novick, 1966). In the case of SPRT being used to classify examinees for the purpose of calibrating ARCH, error of SPRT may be getting compounded in that it is reflected in both the SPRT and then in ARCH post-calibration results. In other words, one misclassification by SPRT could live on to contribute to multiple misclassifications by ARCH post-calibration due to the role SPRT plays in calibrating items used by ARCH post-calibration.

6.2.4 Implications

There are two main implications associated with the findings for RQ2. First, more research is required to determine the specific conditions under which SPRT operates within *a priori* error rates before the ARCH approach can reliably be deployed, as ARCH is heavily dependent on the SPRT testing algorithm. Suggestions for future research on SPRT are outlined in the next chapter. Second, the fact that ARCH post-calibration performed well in terms of classification accuracy in these initial studies (even when SPRT did not)

demonstrates that ARCH represents a promising and robust VL-CCT approach worthy of additional investigation.

6.3 ARCH Efficiency (RQ3)

RQ3 – How efficient is ARCH in comparison to traditionally calibrated SPRT and EXSPRT? – was answered in both study 1 and study 2. As expected, ARCH did not differ from SPRT in terms of mean test lengths or no-decision rates before items had become sufficiently calibrated in both studies. Both ARCH pre-calibration and SPRT methods had significantly longer tests and higher no-decision rates when compared to the EXSPRT method in both studies. After ARCH was able to use calibrated items, ARCH mean test lengths and no-decision rates were significantly smaller than SPRT. However, ARCH had significantly larger mean test lengths compared to EXSPRT across both studies.

6.3.1 Monte Carlo ARCH Evaluation with Historical COM Test Data

In the Monte Carlo simulations with historical COM test data, ARCH pre-calibration test length means and no-decision rates did not differ significantly from SPRT. However, the ARCH post-calibration method provided significant reductions in test length and no-decision rates when compared to SPRT. ARCH post-calibration no-decision rates did not differ significantly from EXSPRT, but ARCH-post calibration mean test lengths were found to be significantly larger than those for EXSPRT.

6.3.2 ARCH Evaluation with New IU Plagiarism Test Examinees

Testing with new examinees associated with a new version of the IU Plagiarism Test revealed that, as expected, ARCH pre-calibration behaved exactly as SPRT with identical test lengths. However, ARCH post-calibration had significantly shorter test lengths and smaller

no-decision rates when compared to SPRT but significantly longer test lengths than EXSPRT.

6.3.3 Discussion of ARCH Efficiency (RQ3)

It is helpful to consider the findings related to the efficiency of ARCH in the context of associated literature, given that use of new adaptive testing techniques to try to reduce test lengths is not new. Wald's (1947) SPRT has been used to reduce test lengths as far back as when Cowden (1947) created an early version of an adaptive classification test. The first example of a VL-CCT from Ferguson (1969) demonstrated that the adaptive version required approximately one-third the number of items as the full test. A study by Linn (1972) provided further support for Green's (1970) assertion that adaptive testing could result in 50 percent shorter tests without compromising accuracy of classifications.

SPRT and ARCH pre-calibration mean test lengths found in the Monte Carlo simulations with COM test historical data are in line with findings from Frick (1989), when based on the same data. Frick found that SPRT had mean test lengths of 18.6 (*SD* = 16.3) for nonmaster decisions and 21.6 (*SD* = 12.9) for master decisions. Recall that mean test lengths in the first study for SPRT were 22.45 (*SD* = 10.79) and for ARCH pre-calibration were 21.98 (*SD* = 9.97). Frick's results and the results from this study represent a 75% reduction in test lengths compared to the entire 85-item test.

EXSPRT and ARCH post-calibration mean test lengths from the first study are also in line with findings from Frick (1992). Frick found that EXSPRT with random item selection had mean tests lengths of 12.82 (*SD* = 9.78), which is very similar to ARCH post-calibration mean test lengths of 13.90 (*SD* = 4.93) and EXSPRT mean test lengths of 13.42 (*SD* = 5.29). Note that the slight differences in test lengths between earlier studies using COM test data

(Frick, 1989; Frick, 1992) can be explained by the fact that these studies involved a single simulated SPRT test and single simulated EXSPRT test for each examinee, whereas, in my study, I conducted 50 simulated SPRT tests for each examinee where items were randomly selected each time. Frick's results and the results from this study for EXSPRT and ARCH post-calibration represent an 84% reduction in test lengths compared to the entire 85 item test. An even larger reduction in test lengths was observed on the IU Plagiarism Test. During IU Plagiarism Testing, ARCH post-calibration had a mean test length of 8.15 (*SD* = 5.03), which represents over a 90% reduction in the number of items compared to the entire item pool. The number of items in the item pool is not specified here for test security reasons.

What is unique about the present studies is that in the ARCH post-calibration method, unlike EXSPRT earlier research (Frick, 1992), items were calibrated without the need for a separate calibration phase prior to starting adaptive testing. The EXSPRT algorithm in Frick's (1992) study was first calibrated with twenty-five responses from nonmasters and twenty-five responses from masters before adaptive testing. The ARCH post-calibration method uses calibration data gathered during live testing via ARCH pre-calibration, which uses the SPRT method.

6.3.4 Implications

The key implication of the results related to the efficiency of ARCH is that substantial reductions in test lengths are achievable in educational contexts without having to calibrate items ahead of time. The efficiency of the ARCH post-calibration method is consistent with both Frick's (1992) examination of the EXSPRT method and Rudner's (2002) analysis of Measurement Decision Theory-based VL-CCT approaches. In both cases, use of item-level

calibration allowed for substantial reductions in test lengths. Item-level calibration data gathered via ARCH pre-calibration/SPRT during live testing enabled ARCH post-calibration to achieve significant reductions in test lengths. However, before ARCH can be considered ready for widespread use, the issues related to the higher than expected error rates observed with SPRT in both studies need to be addressed.

6.4 SPRT False Error Rate Higher Than Expected

The SPRT testing algorithm was found to have a false error rate that was higher than the rate established *a priori* under specific conditions in the first study and under the live testing conditions of the second study. Higher than expected error rates for SPRT are not consistent with earlier research investigating the accuracy of SPRT (Frick, 1989; Frick, 1992; Welch & Frick, 1993; Tao et al., 2008) with two exceptions where SPRT and other VL-CCT approaches were found to be prone to errors when examinees were clustered near cut-scores (Plew, 1989; Frick, 1990). While this finding is does not specifically address any of the three research questions, SPRT is the mechanism by which ARCH classifies examinees as nonmasters or masters during the ARCH pre-calibration phase. Consequently, issues with SPRT could have substantial implications for the ARCH approach to calibration.

6.4.1 Monte Carlo ARCH Evaluation with Historical COM Test Data

During the first study, Monte Carlo ARCH Evaluation with Historical COM Test Data, the SPRT was first calibrated empirically using all the available response data from all nonmasters and masters, which led to a $P(C|M) = .88$ and a $P(C|NM) = .56$. Subsequent simulations of 2,080 SPRT tests using these empirically calibrated values resulted in a SPRT false nonmaster rate of 3.23% versus the *a priori* rate set at 2.5%. When SPRT was manually

set to *P(C|M)* = 0.85 and *P(C|NM)* = 0.60 values, which have been used in previous studies (Frick, 1989; Frick, 1992), false nonmaster rates on an additional 2,080 SPRT based test simulations were below the *a priori* rate of 2.5%. The manually calibrated SPRT made 83 fewer decisions than did the empirically calibrated SPRT method. The mean test length of the tests that applied SPRT with item-bank probabilities set manually was 21.97 items (*SD* = 16.37), which is 5.36 items (32.3%) longer than the results obtained with SPRT using item-bank probabilities set empirically (*M*=16.61, *SD*= 12.28)—but at the expense of more false nonmaster decision errors than expected.

6.4.2 ARCH Evaluation with New IU Plagiarism Test Examinees

During the second study, ARCH Evaluation with New IU Plagiarism Test Examinees, SPRT parameters were set manually to *P(C|M)* = 0.85 and *P(C|NM)* = 0.5. SPRT had a false nonmaster rate of 9.39% on the 1,202 tests during the ARCH pre-calibration phase and 8.71% on the 4,527 tests during the ARCH post-calibration phase, both of which are well above the 5% false nonmaster rate set *a priori*. Unlike the first study, where the examinee total score on the complete 85-item test was used to determine their true mastery state, the second study used the decision of EXSPRT with items calibrated with 50 nonmaster and 50 master responses.

6.4.3 Discussion of SPRT High False Error Rates

SPRT accuracy issues could be the result of coding error or improper SPRT equations. For example, Kingsbury and Weiss (1983) found accuracy problems with SPRT in their study comparing SPRT to IRT-based approaches. However, Frick (1990) suggested that these accuracy issue were the result of using incorrect equations for SPRT. Examination of the code used for SPRT in both the first and second study does not reveal errors, and tests of

the code using input values in previous studies yielded expected outputs. Furthermore, (1) ARCH pre-calibration based on SPRT in the first study yielded very similar results to earlier SPRT simulations done with COM test data (Frick, 1989) when the same *P(C|NM)* and *P(C|M)* settings were used, and (2) coding in the first and second study followed the same coding strategies, albeit with different programming languages (JavaScript versus Python). This suggests that a coding or mathematical error is unlikely to be the reason for SPRT accuracy issues seen in the second study.

A more likely explanation for higher SPRT error rates is that greater *P(C|M)* versus *P(C|NM)* discrimination leads to "rash" and sometimes inaccurate decisions. The fact that SPRT applies item difficulty estimates for all items at the group level (i.e. nonmasters and masters) has been criticized as it ignores the reality that some items are more difficult than others (Weiss & Yoes, 1991; Jacobs-Cassuto, 2005). The greater the difference between *P(C|M)* and *P(C|NM)*, the shorter the tests become, which makes a given SPRT decision dependent on fewer items. SPRT decisions based on a few items may result in increased errors due to differences in item difficulties. For example, on the IU Plagiarism Test there were 604 cases of SPRT making a nonmaster decision after an examinee responded to the first three items incorrectly but in 42 of these 604 cases EXSPRT did not make a decision and continued to administer items. Because EXSPRT uses item level data, it knows when examinees were administered difficult items and does not jump to conclusions when the examinee gets a hard item incorrect.

Another possible explanation for high SPRT error rates involves clustering of examinees near the cut-score. When the shape of distribution of total scores includes examinees bunched around the cut-point, error rates have been shown to be negatively

impacted (Plew, 1989; Frick, 1990). However, since individual examinee responses to all the items on the IU Plagiarism Test are not available, I am unable to determine if there was in fact clustering of examinee total scores around the cut-score.

Another fact to consider is that there were drastically different percentages of nonmasters in the two studies and that this could potentially have impacted false nonmaster error rates. The first study had 28 nonmasters (27%) who responded correctly to less than 72.5% of the items and 76 masters (73%) who responded correctly to at least 72.5% of the 85-item COM test. In the second study, EXSPRT (calibrated with 50 responses from nonmasters and 50 responses from masters to every item) represented the true mastery status. In the second study, 57% of the 5,729 examinees were classified as true nonmasters by EXSPRT, with the remaining 43% being classified as masters. The nonmaster rate in the second study was more than double that of the first study. Rudner (2009) suggested using the proportion reduction in error (PRE) to adjust for the probability that a given algorithm could be right due to chance. With a nonmaster rate in the second study that is double that of the first study, the probability of correctly making a nonmaster decision by chance is also doubled. Further, more nonmasters means more nonmaster decisions and thus more opportunities to make false nonmaster decisions.

It may be the case that SPRT makes nonmastery decision too quickly, which makes it prone to false nonmastery errors. In what may be the first example of a VL-CCT, Ferguson (1969) found that SPRT made nonmaster decision more quickly than master decisions, and similar findings have come from subsequent SPRT research (Frick, 1989; Frick, 1992). However, in these studies SPRT error rates were not found to be problematic.

6.4.4 Implications

Higher-than-expected SPRT false nonmaster rates under some conditions is troublesome given the core role that SPRT plays in the ARCH approach. SPRT is the testing algorithm that ARCH uses to make classification decisions while it is still gathering calibration information about items, so issues with SPRT accuracy become issues with ARCH accuracy, as was demonstrated in the second study with new IU Plagiarism test examinees.

Equally problematic is the fact that the classification decisions made by ARCH pre-calibration/SPRT about a specific examinee drives how items are calibrated. If SPRT/ARCH pre-calibration makes a nonmaster decision for an examinee, all nonmaster-related item calibration data (i.e., number of correct and incorrect responses by nonmasters to a given item) for all the items to which the examinee responded get updated accordingly. Similarly, all master-related item calibration for all the items to which the examinee responded get updated when SPRT/ARCH pre-calibration makes a master decision.

False nonmaster decisions mean that nonmaster calibration data are incorrectly being updated and master calibration data are incorrectly not getting updated. Consequently, it is reasonable to posit that inaccurate item calibration data gathered during ARCH pre-calibration could contribute to increasing the error rate of ARCH post-calibration, which can be viewed as a compounding of the error rate. However, evidence from the IU Plagiarism Test study where SPRT/ARCH pre-calibration had a higher than expected false error rate did not show a corresponding higher than expected error rate with ARCH post-calibration. Results from the Monte Carlo simulations with historical COM test data included ARCH post-calibration false nonmaster error rates that were significantly higher than those for

SPRT – the exact opposite of what was observed in the IU Plagiarism Test – so it is difficult to draw any clear conclusions regarding the relationship between false nonmaster error rates between SPRT/ARCH pre-calibration and ARCH post-calibration.

Given that SPRT false error rates may be associated with greater differences between *P(C|M)* and *P(C|NM)* settings for SPRT, it may be advisable to restrict SPRT *P(C|M)* and *P(C|NM)* settings such that differences are in line with or smaller than settings (i.e. *P(C|M)* = 0.85 and *P(C|NM)* = 0.60) that have been shown to result in SPRT performing within *a priori* error rates (Frick, 1989; Frick, 1992; Welch & Frick 1993). Furthermore, care must be taken to avoid using SPRT in cases where examinees are likely to be clustered around the cut-score since this can contribute to higher than expected error rates (Plew, 1989; Frick, 1990). In addition, *a priori* false nonmaster and false master error rates can be adjusted to lower, more conservative, values as a means of reducing error rates. The 0.25 difference between *P(C|M)* = 0.85 and *P(C|NM)* = 0.60 can be achieved with other settings (e.g., *P(C|M)* = 0.90 and *P(C|NM)* = 0.65, *P(C|M)* = 0.70 and *P(C|NM)* = 0.45), so SPRT decision can reflect the nature of the nonmaster versus master decision that needs to be made. The likely consequence of minimizing the difference between *P(C|NM)* and *P(C|M)* is increasing test lengths, and the appropriateness of this tradeoff must be evaluated by test administrators in a given assessment context.

6.5 Viability of ARCH in Real-World Contexts

While ARCH does not require a separate calibration phase before adaptive testing can begin, large numbers of test administrations are required before ARCH shifts to the most efficient testing approaches. In the first study using historical COM test data, around 700 test administrations were required in repeated Monte Carlo simulations before the 85 items were

sufficiently calibrated to enable ARCH to shift to adaptive testing using EXSPRT based methods. With a larger item pool, over 4,600 test administrations were required before ARCH could switch to the most efficient adaptive testing methods on the IU Plagiarism Test. The main drawback associated with IRT (Thompson, 2007; Frick, 1992) is that a large calibration sample is required to establishing item parameter estimates. Given that hundreds or thousands of test administrations were required before ARCH could switch to the most efficient adaptive testing methods in both studies, it would seem that ARCH is similar to IRT in requiring a large calibration sample, however, there are important differences.

ARCH calibration involves many examinees answering a small percentage of the total number of items in the item pool under actual testing conditions, whereas traditional calibration of EXSPRT involves having at least 25 nonmasters and 25 masters respond to *all the items in the item bank*. Large numbers of test administrations are typically not a problem in massively open online contexts. For example, in the second study with the IU Plagiarism Test only a 25-hour period was required before item ARCH calibration criteria had been reached and more efficient testing algorithms could be deployed. However, the ARCH approach may not be viable in situations where the number of potential examinees is small. The reader should note that the item pool on the IU Plagiarism Test is very large, and many examinees took the test who could not recognize plagiarism. The ratio of nonmasters to masters was about 4 to 1. Thus it took a large number of test administrations to identify enough masters to satisfy the EXSPRT minimum for every item. This was not only due to the relatively small proportion of mastery decisions, but also due to the fact that tests were relatively short (8.55 items on average were sampled from the very large pool), so it took a very large number of test administrations before *each* item was viewed by enough masters.

If a much smaller item pool were used, and test items were easier overall, then the ARCH post-calibration method could be employed much sooner than observed in the second study here.

The ARCH approach to item calibration has additional advantages over traditional EXSPRT calibration with a set calibration sample. With ARCH, not a single examinee sees all the items in the pool, but with traditional EXSPRT calibration at least 50 individuals would see the entire item pool, which presents considerable test security issues. Another advantage for the ARCH approach to calibration is that item calibration data are generated during live testing, which means examinees are motivated to perform well. Examinee performance during low or no stakes item calibration phases has been raised as an issue that potentially impacts the quality of item calibration data collected (Wise & DeMars, 2006 in Makransky, 2010).

Ultimately, the quality and reliability of the ARCH testing approach, like all testing approaches, requires that the relationships between the observed score/classification, the true score/classification, and error be examined and understood (Allan & Yen, 2002). Findings from both studies suggest that ARCH is a promising new VL-CCT approach. However, further research is required to more fully understand the relationship between classification errors made during SPRT/ARCH pre-calibration and their impact on ARCH post-calibration.

6.6 Practical Implications

The ARCH approach to item calibration and computer adaptive testing is well suited to specific types of assessment contexts. The ARCH approach enables efficient and accurate

criterion-referenced assessment without the arduous item-calibration requirements associated with IRT based approaches. Consequently, ARCH is well suited to criterion-referenced assessment contexts where (1) the resources needed for IRT-based item calibration methods are not available or not justifiable; and (2) the necessary technology for CAT is in place but the resources required for IRT-based testing approaches are not. The following presents several assessment contexts where the ARCH approach to item calibration is particularly applicable followed by a brief discussion of when ARCH is not advisable.

A general example of a context where ARCH would be well suited is associated with enabling the information-age learning management system (LMS) that Reigeluth and colleagues (2008) have proposed. Central to an information-age LMS is the ability to make classification decisions about what an individual learner knows in order to select appropriate learning experiences that the learner should next attempt. Rather than the standard practice of giving an entire *group* the same test at the end of a given unit of instruction before moving onto the next unit, in an information-age LMS *individual* learners are assessed several times to facilitate the selection of an appropriate learning experience and then to ensure they have mastered the associated learning outcomes. Learners do not proceed until they have mastered the target competencies so a learner may need to be assessed multiple times before mastery is attained.

In the context of an information-age LMS, ARCH is uniquely suited to provide efficient and accurate mastery testing. IRT based methods would require too many resources for it to be feasible to create the large number of mastery focused tested that address numerous learning outcomes that would be assessed in an information-age LMS. In an information-age LMS it is critical to ensure that time spent conducting assessments takes a

minimal amount of time away from learning time so CAT methods would be important as a means of reducing testing time. The fact that many learners would need to take tests multiple times before attaining a mastery score would increase item exposure and make it important to be able to quickly create and calibrate new items to reduce the chances that the same learners would be administered the same items during multiple attempts to achieve mastery on the tests.

A more specific example a context that is well suited to the ARCH approach is assessment of the effectiveness of classroom technology integration efforts. When students are learning in technology-enhanced environments it makes sense to apply technology enhanced assessment such as CAT. In some cases, evaluation of technology-oriented skills may not be possibly via traditional pencil and paper tests. Furthermore, given the pace at which technology has been changing, the time it would take to create and calibrate a pool of items using IRT-based methods would represent an undesirable delay that could hinder the applicability and longevity of technology oriented items developed. Items that reference specific technologies have a short shelf life, so reductions in the calibration burden provided by ARCH are helpful in quickly putting new items into operation.

For example, recent increased adoption of Google Chromebooks in classroom settings is replacing Apple and Microsoft based computing devices (Winkler, 2014). An item that evaluates an examinee's knowledge of how to share a file with a peer using a Google Chromebook that is relevant today is likely to become outdated as new methods of sharing pictures via a Chromebook become available, or Chromebooks themselves become outdated technology.

ARCH can be helpful in educational contexts where large-scale standardized assessments are also deployed. Large-scale standardized tests are typically norm-referenced and provide summative percentile rank examinee test results for a given generalized construct that is often more useful to researchers, administrators, and policy makers than to students, parents, and educators (Hickey, 2006). VL-CCT that apply ARCH calibration methods make classification decisions (e.g. mastery versus nonmastery) with respect to specific learning outcomes and, as such, reveal more actionable information about the state of a learner's knowledge that educators, parents, and the students themselves can use. The fact that ARCH calibrates items more quickly than IRT based methods means that the associated tests can be deployed in lower-stakes contexts where test security is not as high, such as in classrooms or via a course space in an LMS.

Finally, as was demonstrated through the use of ARCH in the context of the IU plagiarism test, ARCH is particularly useful in massively open online assessment contexts (MOOCs). MOOCs have specific attributes that present challenges for assessment. Their massively open nature limits the control that test administrators have over the test environment and makes it difficult to hinder cheating. ARCH based VL-CCTs help to hinder cheating in MOOCs through use of random item selection, and also help to ease development of a large calibrated item bank.

The ARCH approach and other computer-based testing approaches are not well suited in other types of assessment contexts. ARCH is not appropriate if test items cannot be easily administered or evaluated using computing resources. Administering a valid test item where the physical actions performed by the examinee constitute the response (e.g. performing first aid to address particular injuries presented by a simulated victim) would be challenging, if

not impossible, in a computer-based testing context. Automatically evaluating the correctness of certain types of response such as an essay-style written responses to open-ended ill-structured problems (e.g. Shin, Jonassen, & McGee, 2003) would be challenging using only computing resources. Although advances in artificial intelligence such as deep learning are pushing the boundaries of what computers can do, current computers are nonetheless limited with respect to understanding meaning of natural language and observing human actions (e.g., see Frick, 1997).

If the construct being assessed is only applicable to a relatively small number of potential examinees then ARCH should likely not be used since it would likely take a very long time for items to be calibrated.

Finally, ARCH is not appropriate for norm-referenced testing, since ARCH is an approach to criterion-referenced testing. However, if U.S. school systems were to adopt a criterion-referenced approach, instead of massive statewide assessments being conducted annually at the same time, CRT's could instead be administered individually as needed for students throughout the school year. Students would be taking appropriate CRTs for standards on which they are ready to be assessed. Such individualization and staggering of computer-based tests would help alleviate possible problems associated with large numbers of students take the same norm-referenced test at the same time which can swamp web servers and network capacities for the testing agencies, cause delays between test items, or worse, result in computer crashes where student data is lost (Rabinowitz & Brandt, 2001). Using CRTs that are staggered throughout the school year would help alleviate such "traffic jams" that occur when literally tens of thousands of students are all tested at the same time. This does not require abandonment of standards for student achievement, rather a different

approach to assessment of those standards. For such a vision of such individualized assessment that was proposed some time ago, see Frick (1991, https://www.indiana.edu/~tedfrick/fastback/fastback326.html#student-content , “What If”) and Frick (1990, p. 480).

CHAPTER VII. SUMMARY, LIMITATIONS, AND FUTURE RESEARCH

7.1 Summary

Use of Computer Adaptive Testing (CAT) has tremendous potential in educational contexts to quickly and accurately assesses learner knowledge (Frick, 1989; Frick, 1992; Rudner, 2009). Massively Open Online Courses in particular require a reconceptualization of traditional assessment practices (DeBoer et al., 2014) with CAT uniquely positioned to efficiently assess large numbers of learners. However, use of CAT is hindered by arduous item calibration requirements, in some cases involving thousands of examinees (Weiss & Kingsbury, 1984; Welch & Frick, 1993) that have limited the use of CAT to large-scale, high-stakes, and/or highly profitable contexts. Highly efficient and accurate Variable-Length Computerized Classification Testing (VL-CCT) methods requiring limited item calibration are available (Rudner 2009; Thompson, 2007; Frick, 1992), but minimal research has been conducted on their calibration requirements (Frick, 1992).

The purpose of my research was to develop and evaluate an innovative item calibration and VL-CCT method, Automatic Racing Calibration Heuristics (ARCH), that includes specific item calibration guidelines and an approach that shifts to a more efficient VL-CCT approach, EXSPRT, as items are sufficiently calibrated during live testing. My research involved two studies.

The first study, Monte Carlo ARCH operationalization and evaluation with historical COM test data, addressed three research questions: (RQ1) When is an item sufficiently calibrated? (RQ2) How accurate is ARCH? and (RQ3) How efficient is ARCH? Data for the first study were drawn from 104 examinee responses to an 85-item multiple-choice test on how computers work that has been the subject of previous studies (Frick, 1989; Frick, 1992).

Thousands of Monte Carlo simulations were conducted to examine how specific ARCH calibration criteria and associated thresholds impacted adaptive test accuracy and efficiency and, ultimately, to establish the calibration criteria and thresholds that answer RQ1. ARCH pre-calibration and post-calibration was then compared to the SPRT and EXSPRT testing algorithms via 50 sets of Monte Carlo simulations based on response data from each of the 104 examinees. ARCH error rates were compared to *a priori* error rates through chi-squared goodness-of-fit testing using exact nonasymptotic methods. ARCH error rates, mean test lengths, and no-decision rates were compared to those of SPRT and EXSPRT through Friedman Tests and post hoc analysis using Wilcoxon signed-rank testing.

The set of calibration criteria and associated thresholds that provide an answer to RQ1 regarding calibration sufficiency are: (1) beta *SD* of $P(C_i|M) \leq .078$; (2) beta *SD* of $P(C_i|NM) \leq .078$; (3) *Beta Difference Index* of difference between $P(C_i|M)$ and $P(C_i|NM) \geq 15$; and (4) the percent of items calibrated $\geq 83\%$. With respect to RQ2, the first study found that ARCH did not deviate significantly from *a priori* error rates in the vast majority of cases. ARCH error rates did not differ significantly from SPRT and EXSPRT, with the exception of ARCH post-calibration committing significantly more false nonmaster errors than SPRT. In terms of test efficiency (RQ3), ARCH post-calibration provided significant reductions in test length and no-decision rates when compared to SPRT. ARCH post-calibration did not differ from EXSPRT in terms of no-decision rates but the marginally longer mean test lengths of ARCH-post calibration were found to be significantly different from those of EXSPRT.

The second study, ARCH evaluation with new IU plagiarism test examinees, applied the findings of RQ1 from the first study and answered (RQ2) and (RQ3) using data gathered

in December 2013 and January 2014 from the massively open online context of a new version of the Indiana University Plagiarism Test, which I developed, that included a large item pool. Test results and response data from nearly 15,000 unique participants were used to calibrate all the items in the pool with 50 responses from nonmasters and 50 responses from masters for subsequent use with EXSPRT-based testing.

After EXSPRT calibration was complete, test results and response data were collected from 5,729 examinees for SPRT, ARCH, and EXSPRT. The first 1,202 examinees took the ARCH pre-calibration version that mimicked SPRT and gathered calibration data, with the remaining 4,527 taking the ARCH post-calibration version that used calibration information gathered during live testing and a modified version of EXSPRT to render classification decisions. Again, ARCH error rates were compared to *a priori* error rates through chi-squared goodness-of-fit testing using exact non-asymptotic methods. ARCH error rates, mean test lengths, and no-decision rates were compared to those of SPRT through Friedman Tests and post hoc analysis using Wilcoxon signed-rank testing.

In terms of ARCH accuracy (RQ2), the second study found that ARCH did deviate significantly from *a priori* error rates, with ARCH pre-calibration having higher error rates than expected and ARCH post-calibration having lower error rates than expected. ARCH pre-calibration error rates were exactly the same as those for SPRT, since ARCH pre-calibration mimics SPRT. ARCH post-calibration had significantly lower error rates than SPRT. With respect to efficiency (RQ3), ARCH post-calibration provided significant reductions in test length and no-decision rates when compared to SPRT. ARCH post-calibration had marginally longer mean test lengths than EXSPRT and this difference of approximately one item was found to be statistically significant.

Major findings of these two studies include a refined version of ARCH, the development of new item calibration/quality statistics, and establishment of item calibration statistic threshold values that led to efficient and accurate VL-CCT in multiple contexts. While not the focus of my research, the SPRT testing algorithm was found to be prone to higher than expected false nonmaster error rates under specific simulation conditions and during live testing. Results from both studies led to the establishment of specific criteria indicating when items are sufficiently calibrated and suggest that ARCH did enable accurate and efficient VL-CCT without the need for a separate item calibration phase.

7.2 Limitations

This research is the first of its kind. These were the very first two studies of the ARCH calibration and adaptive testing method. More studies are needed before generalizations about the ARCH approach can be made and ARCH can be considered an established VL-CCT method.

Moreover, ARCH introduces and depends on new item calibration and quality statistics such as the beta *SD* and the *Beta Difference Index*. While these new item calibration and quality criteria performed well in both studies to indicate when items had been sufficiently calibrated, they have not yet been subject to scrutiny from the broader educational assessment research community.

The fact that the two contexts in which ARCH was applied are extremely different may be considered another limitation. The historical data gathered from the COM test took place in a traditional face-to-face context involving 104 volunteer examinees answering 85 multiple-choice questions at VT-240 computer terminals in a computer lab under proctored

conditions in the 1980s. The web-based IU Plagiarism Test occurred in an unproctored, massively open, online context in late 2013 and early 2014. Thousands of examinees responded to multiple-choice questions randomly selected from a large item pool, with the goal of earning a certificate that confirmed they could recognize plagiarism. Differences in the contexts, numbers of examinees, examinee motivations, and timeframes spanning four decades present challenges when comparing the two contexts.

Another limitation is that decisions made regarding the IU Plagiarism Test study had to be balanced against practical consequences that the decisions would have on this heavily used online resource. For example, use of the same SPRT probability values for *P(C|NM)* and *P(C|M)* and *a priori* error rates as those used in earlier COM test studies (Frick, 1989; Frick, 1992) would have been advantageous from a research perspective to enable more direct comparisons. However, the new adaptive version of the IU Plagiarism Test was found to be very difficult for many examinees, and compromises were made to keep test lengths more reasonable. Consequently, *a priori* error rates for false nonmastery and false mastery decisions were relaxed; and SPRT probabilities were adjusted for practical rather than research reasons. This had the desired effect of shorter CCT Plagiarism tests, but at the cost of higher-than-expected false nonmastery decision error rates.

Finally, using the EXSPRT decision outcomes to indicate true examinee mastery status on the IU Plagiarism test is a limitation. VL-CCT based on Classical Test Theory depends on the ability to distinguish between classification groups, ideally via a method independent of the specific test (e.g., a separate test, expert judgment, etc.) (Frick, 1992; Rudner, 2002; Thompson, 2007). However, in the context of the IU Plagiarism Test there

was no independent way of determining mastery status efficiently for the volume of examinees involved.

7.3 Suggestions for Future Research

Suggestions for future research outlined below stem from several of the issues with the ARCH approach outlined in the discussion chapter. Higher-than-expected false nonmaster rates associated with the SPRT testing algorithm warrant further investigation given the central role that SPRT plays in ARCH. Monte Carlo simulation studies could be conducted to establish the conditions under which SPRT operates within *a priori* error rates and when it does not. For example, findings from both studies suggest that there may be a relationship between the high error rates observed with SPRT and the difference between *P(C|M)* and *P(C|NM)* probability values set for SPRT. The potential relationship between false error rates and *P(C|M)* and *P(C|NM)* probability values could be examined by conducting Monte Carlo simulations where the *P(C|M)* and *P(C|NM)* probability rates are systematically adjusted in order to examine the resulting impact on observed error rates.

Investigation of methods for reducing the number of misclassified examinees that contribute to item calibration data used by ARCH post-calibration provides another potential line of research. Given that the vast majority of SPRT errors in the Monte Carlo simulations with COM test data occurred with examinees near the cut score, a future study could examine ways of identifying borderline examinees so that their data could be excluded from ARCH item calibration. Incorrect classification of those examinees whose overall test scores would be in Wald's zone of indifference (between masters and nonmasters) may have contributed to higher ARCH post-calibration error rates observed in the first study. One strategy would be to have ARCH pre-calibration (SPRT) use three classification categories – nonmaster, zone

of indifference, and master – so that those examinees likely near the border in the zone of indifference are not used to calibrate items. Another strategy would be to use stricter decision error rates with SPRT during ARCH pre-calibration to limit the number of false nonmaster and false master decisions; however, stricter error rates would likely come at the cost of increasing average test lengths.

REFERENCES

Allen, M. J., & Yen, W. M. (2001). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press Inc.

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The 1996 NAEP Technical Report.* ERIC. Retrieved from http://eric.ed.gov/?id=ED432620

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, *12*(1), 137–144.

Barrett, A. F. (2010). Measuring TPK component of TPACK: An Alternative to Self-Assessment. Presented at the American Educational Research Association Conference, Denver, Colorado.

Barrett, A. F., & Howard, C. D. (2010). Validity of Computer Mediated Formative Peer Assessment: Pre-service Teacher's Comments in Asynchronous CMC. Presented at the American Educational Research Association Conference, Denver, Colorado.

Binet, A., & Simon, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'année Psychologique*, *11*(1), 191–244.

Black, P., & Wiliam, D. (1998). Inside the Black Box: Raising Standards through Classroom Assessment. *Phi Delta Kappan*, *80*(2).

Black, P. J. (1993). Formative and summative assessment by teachers. *Studies in Science Education*, *21*(1), 49–97.

Brush, T., Glazewski, K. D., & Hew, K. F. (2008). Development of an Instrument to Measure Preservice Teachers' Technology Skills, Technology Beliefs, and Technology Barriers. *Computers in the Schools*, *25*(1), 112-125.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University Testing Services and The Department of Instructional Science. Retrieved from http://testing.byu.edu/faculty/handbooks.asp

Chang, Y., & Lu, H. (2010). Online Calibration Via Variable Length Computerized Adaptive Testing. *Psychometrika*, *75*(1), 140-157. doi:10.1007/s11336-009-9133-0

Christensen, C., Johnson, C. W., & Horn, M. B. (2008). *Disrupting Class: How Disruptive Innovation Will Change the Way the World Learns* (1st ed.). McGraw-Hill.

Cohen, J. (1988). *Statistical power for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155 - 159. doi:10.1037/0033-2909.112.1.155

Collins, A., & Halverson, R. (2009). *Rethinking Education in the Age of Technology: The Digital Revolution and Schooling in America (Technology, Education--Connections (Tec))* (1st ed.). Teachers College Press.

Colorado State Department of Education (2000). Colorado Student Assessment Program (CSAP), technical report, grade 5 mathematics. Available online: http://www.cde.state.co.us/cdeassess/download/pdf/as_csaptech5math99.pdf

Cowden, D. J. (1946). An application of sequential sampling to testing students. Journal of the American Statistical Association, 41(236), 547–556.

Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions. Urbana, IL: University of Illinois Press.

Crooks, T. (1988). Impact of classroom assessment on students. *Review of Educational Research*, *55*(4), 438–481.

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "Course" Reconceptualizing Educational Variables for Massive Open Online Courses. *Educational Researcher*, *43*(2), 74–84. doi:10.3102/0013189X14523038

Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment*, *4*(1), 37–73.

Ebel, R. L. (1972). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Eggen, T., & Straetmans, G. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological measurement*, *60*(5), 713.

Eggen, T. J. H. M. (1999). Item Selection in Adaptive Testing with the Sequential Probability Ratio Test. *Applied Psychological Measurement*, *23*(3), 249-261. doi:10.1177/01466219922031365

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing (ed J. Weiss)*. Presented at the GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. doi:10.1177/0013164498058003001

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175.

Ferguson, R. L. (1969). *Computer-Assisted Criterion-Referenced Measurement.* (No. Report WP-41). Pittsburgh: Pittsburgh University, Learning Research and Development Center.

Frick, T. W. (1989). Bayesian adaptation during computer-based tests and computer-guided practice exercises. *Journal of Educational Computing Research*, *5*(1), 89-114.

Frick, T. W. (1990). A comparison of three decision models for adapting the length of computer-based mastery tests. *Journal of Educational Computing Research*, *6*(4), 479–513.

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research*, *8*(2), 187-213.

Frick, T. W. (1997). Artificial Tutoring Systems: What Computers Can and Can't Know. *Journal of Educational Computing Research*, *16*(2), 107-24.

Frick, T. W. & Thompson, K. R. (2008). Predicting education system outcomes: A scientific approach. In M. Orey, V. McClendon, & R. Branch (Eds.), *Educational Media and Technology Yearbook* (Vol. 33), 62-76.

Gardner, D. P., Larsen, Y. W., Baker, W., & Campbell, A. (1983). A nation at risk: The imperative for educational reform. *Washington, DC: US Government Printing Office*. Retrieved from http://mathcurriculumcenter.org/PDFS/CCM/summaries/NationAtRisk.pdf

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*(8), 519.

Green Jr, B. F. (1970). Comments on Tailored Testing. *Computer-Assisted Instruction, Testing, and Guidance*, 184.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*(3), 159-170. doi:10.1111/j.1745-3984.1973.tb00793.x

Hambleton, R. K., & Xing, D. (2006). Optimal and Nonoptimal Computer-Based Test Designs for Making Pass–Fail Decisions. *Applied Measurement in Education*, *19*(3), 221. doi:10.1207/s15324818ame1903_4

Hew, K. F., & Brush, T. (2007). Integrating technology into K-12 teaching and learning: current knowledge gaps and recommendations for future research. *Educational Technology Research and Development*, *55*(3), 223-252.

Hickey, D. T., Zuiker, S. J., Taasoobshirazi, G., Schafer, N. J., & Michael, M. A. (2006). Balancing varied assessment functions to attain systemic validity: Three is the magic number. *Studies in Educational Evaluation*, *32*(3), 180-201.

Hutt, M. L. (1947). A clinical study of "consecutive" and "adaptive" testing with the revised Stanford-Binet. *Journal of Consulting Psychology*, *11*(2), 93–103.

Hylén, J., & Schuller, T. (2007). Giving knowledge for free. *OECD Observer*, *263*, 21–22.

Jacobs-Cassuto, M. S. (2005). *A Comparison of Adaptive Mastery Testing Using Testlets With the 3-Parameter Logistic Model*. University of Minnesota, St. Paul, MN.

Kelley, T., L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, *30*(1), 17–24. doi:10.1037/h0057123

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. *New horizons in testing: Latent trait test theory and computerized adaptive testing*, *44*.

Lee, P. M. (2004). *Bayesian statistics*. New York, NY, USA: John Wiley & Sons Inc.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*(4), 367.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, *29*(1), 129.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1972). Sequential testing for dichotomous decisions. *Educational and Psychological Measurement*, *32*(1), 85.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores. Retrieved from http://doi.apa.org/psycinfo/1968-35040-000

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo Comparison of Item and Person Statistics Based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement*, *62*(6), 921–943. doi:10.1177/0013164402238082

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*(2), 99–120.

Macready, G. B., & Mitchell Dayton, C. (1992). The application of latent class models in adaptive testing. *Psychometrika*, *57*(1), 71–88.

Makransky, A. G. (2008). Computer adaptive testing: An introduction to the practial considerations related to developing a computer adaptive test for occupational testing. *K. Benny, Metodologiske Indblik Og Udsyn*, 197–223.

McCombs, B. L., & Whisler, J. S. (1997). *The learner-centered classroom and school: Strategies for increasing student motivation and achievement*. Jossey-Bass Inc Pub.

Moonan, W. J. (1950). Some empirical aspects of the sequential analysis technique as applied to an achievement examination. *The Journal of Experimental Educational*, 195–207.

Mourshed, M., Chijioke, C., & Barber, M. (2011). How the worlds most improved school systems keep getting better. *Educational Studies*, (1), 7–25.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18.

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

Parshall, C. G., Davey, T., & Pashley, P. (2000). Innovative item types for computerized testing. *Computerized adaptive testing: Theory and practice*, 129–148.

Parshall, C. G., & Harmes, J. C. (2007). Designing Templates Based on a Taxonomy of Innovative Items.

Plew, G. T. (1989). *An empirical investigation of major adaptive testing methodologies and an expert systems approach*. Indiana University, United States -- Indiana.

Popham, W. J. (2003). *Test Better, Teach Better: The Instructional Role of Assessment*. Association for Supervision and Curriculum Development.

Powell, Z. E. (1992). Test Anxiety and Test Performance Under Computerized Adaptive Testing Methods. Retrieved from http://eric.ed.gov/?id=ED344910

Rabinowitz, S., & Brandt, T. (2001). Computer-Based Assessment: Can It Deliver on Its Promise? Knowledge Brief. Retrieved from http://eric.ed.gov/?id=ED462447

Reigeluth, C. M., Watson, W. R., Watson, S. L., Dutta, P., Chen, Z. C., & Powell, N. D. (2008). Roles for technology in the information-age paradigm of education: learning management systems. *Educational technology*, *48*(6), 32-39.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, 237–255.

Reynolds, C. R., Livingston, R. B., Willson, V. L., & Willson, V. (2010). *Measurement and assessment in education*. Pearson Education International. Retrieved from http://library.mpib-berlin.mpg.de/toc/z2007_216.pdf

Rudner, L. M. (1998). An on-line, interactive, computer adaptive testing tutorial. *ERIC Clearinghouse on Assessment and Evaluation*.

Rudner, L. M. (2002a). An examination of decision-theory adaptive testing procedures. *Proceedings of American Educational Research Association*, 437–446.

Rudner, L. M. (2002b). *Measurement decision theory*. Retrieved from http://ericae.net/mdt/

Rudner, L. M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research & Evaluation*, *14*(8). Retrieved from http://pareonline.net/getvn.asp?v=14&n=8.

Rudner, L. M. (2010). Implementing the Graduate Management Admission Test Computerized Adaptive Test. *Elements of Adaptive Testing*, 151–165.

Schmidt, D., Baran, E., Thompson, A., Koehler, M., Punya, M., & Shin, T. (2009). Examining Preservice Teachers' Development of Technological Pedagogical Content Knowledge in an Introductory Instructional Technology Course.

Schmidt, D., Baran, E., Thompson, A., Koehler, M., Punya, M., & Shin, T. (2009). Examining Preservice Teachers' Development of Technological Pedagogical Content Knowledge in an

Introductory Instructional Technology Course. In *Proceedings of Society for Information Technology & Teacher Education International Conference 2009* (Vol. 2009, pp. 4145-4151). Presented at the Society for Information Technology & Teacher Education International Conference 2009, Chesapeake, VA: AACE. Retrieved from http://editlib.org/p/31308

Schmitt, S. A. (1969). *Measuring uncertainty: An elementary introduction to Bayesian statistics*. Reading, MA: Addison-Wesley.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement*, *16*(1), 65.

Shrock, S. A., & Coscarelli, W. C. (2007). *Criterion-referenced test development: technical and legal guidelines for corporate training*. Hoboken, NJ: John Wiley and Sons.

Spector, J. M. (2014). Remarks on MOOCS and Mini-MOOCS. *Educational Technology Research and Development*, *62*(3), 385–392. doi:10.1007/s11423-014-9339-4

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21*(4), 405.

Tao, Y. H., Wu, Y. L., & Chang, H. Y. (2008). A Practical Computer Adaptive Testing Model for Small-Scale Scenarios. *Educational Technology & Society*, *11*(3), 16.

Thompson, N. A. (2007). A Practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation*, *12*(1). Retrieved from http://pareonline.net/getvn.asp?v=12&n=1

Triantafillou, E., Georgiadou, E., & Economides, A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education*, *50*(4), 1319–1330.

Vos, H. J. (1997). A simultaneous approach to optimizing treatment assignments with mastery scores. *Multivariate Behavioral Research*, *32*(4), 403–433.

Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *24*(3), 271.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In *Computerized adaptive testing: A primer* (Vol. 2). Mahwah, NJ: Lawrence Erlbaum Associates.

Wald, A. (1946). Differentiation under the expectation sign in the fundamental identity of sequential analysis. *The Annals of Mathematical Statistics*, 493–497.

Wald, A. (1947). *Sequential analysis.* New York, NY: Wiley.

Wald, A. (1973). *Sequential analysis*. Courier Corporation. Retrieved from http://books.google.ca/books?hl=en&lr=&id=zXqPAQAAQBAJ&oi=fnd&pg=PP1&dq=Wald+Sequential+Analysis&ots=lXJcSIdsCI&sig=9OSpq6x1GdocQOQgJ3ekFDAU2tg

Watson, S. L., & Reigeluth, C. M. (2008). The Learner-Centered Paradigm of Education. *Educational Technology Magazine: The Magazine for Managers of Change in Education*, *48*(5), 7.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53*(6), 774–789.

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. *Advances in educational and psychological testing: theory and applications*, 69–95.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*, *21*(4), 361-375.

Welch, E. R., & Frick, T. W. (1993). Computerized adaptive testing in instructional settings. *Educational Technology Research and Development*, *41*(3), 47-62.

Welch, R. E. (1997). *Computerized adaptive testing methodologies under real-time testing conditions: A study of efficacy and efficiency*. Indiana University, Bloomington, IN.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116.

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, *64*(1), 5.

Xu, T., & Stone, C. A. (2012). Using IRT Trait Estimates Versus Summated Scores in Predicting Outcomes. *Educational and Psychological Measurement*, *72*(3), 453–468. doi:10.1177/0013164411419846

Yang, X., Poggio, J. C., & Glasnapp, D. R. (2006). Effects of estimation bias on multiple-category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement*, *66*(4), 545.

APPENDIX A: Abbreviations

Abbreviation	*Description*
ARCH	Automatic Racing Calibration Heuristics
α	Probability of making a type I error
β	Probability of making a type II error
beta (.6 | 5, 10)	The ordinate of the BETA density for p = .6 when the parameters are s = 5 and f = 10
beta (* | 5, 10)	Probability density function when the parameters are s = 5 and f = 10
BETA (.6 | 5, 10)	Probability of a BETA variable with s = 5 and f = 10 is less than or equal to .6
BETA (* | 5, 10)	Cumulative distribution function with s = 5 and f = 10 is less than or equal to .6
CRT	Criterion-Referenced Testing
CTT	Classical Test Theory
E (beta * | 5, 10)	The expected value (mean) of a beta variable with parameters s=5 and f=10. = (s + 1)/(s + f + 2)
EXSPRT-R	Expert systems reasoning applied to Sequential Probability Ratio Test (SPRT) that uses random item selection
EXSPRT-I	Expert systems reasoning applied to Sequential Probability Ratio Test (SPRT) that uses intelligent item selection
f	Failures

HDRW	Highest Density Region Width
IRT	Item Response Theory
M-EXSPRT-R	Measured application of expert systems reasoning applied to Sequential Probability Ratio Test (SPRT) that uses random item selection
p	Probability
pd	Probability Density
P	Proportion of Successes
s	Successes
SPRT	Sequential Probability Ratio Test
VL-CCT	Variable-Length Computer Classification Tests
XML	Extensible Markup Language

APPENDIX B: Initial Ten Item Plagiarism Test

Please note: If the student version contains BOTH word-for-word and paraphrasing plagiarism, you should check word-for-word.

Item 1

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version**
The concept of *systems* is really quite simple. The basic idea is that a system has parts that fit together to make a whole; but where it gets complicated - and interesting - is how those parts are connected or related to each other. There are many kinds of systems: government systems, health systems, military systems, business systems, and educational systems, to name a few. References: Frick, T. (1991). *Restructuring education through technology.* Bloomington, IN: Phi Delta Kappa Educational Foundation.	Systems, including both business systems, and educational systems, are actually very simple. The main idea is that systems have parts that fit together to make a whole. What is interesting is how those parts are connected together.

Which of the following is true for the Student Version above?

- o **Word-for-Word plagiarism**
- o **Paraphrasing plagiarism**
- o **This is not plagiarism**

Item 2

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version**

There is a design methodology called rapid prototyping, which has been used successfully in software engineering. Given similarities between software design and instructional design, we argue that rapid prototyping is a viable method for instructional design, especially for computer-based instruction. References: Tripp, S. D., & Bichelmeyer, B. A. (1990). Rapid prototyping: An alternative instructional design strategy. *Educational Technology Research and Development, 38*(1), 31-44.	Rapid prototyping could be an advantageous methodology for developing innovative computer-based instruction (Tripp & Bichelmeyer, 1990). References: Tripp, S. D., & Bichelmeyer, B. A. (1990). Rapid prototyping: An alternative instructional design strategy. *Educational Technology Research and Development, 38*(1), 31-44.

Which of the following is true for the Student Version above?

- o **Word-for-Word plagiarism**
- o **Paraphrasing plagiarism**
- o **This is not plagiarism**

Item 3

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version**
The study of learning derives from essentially two sources. Because learning involves the acquisition of knowledge, the first concerns the nature of knowledge and how we come to know things.... The second source in which modern learning theory is rooted concerns the nature and representation of mental life. References: Driscoll, M. P. (2000). *Psychology of learning for instruction* (2nd Ed.). Needham Heights, MA: Allyn & Bacon.	The study of learning derives from essentially two sources. The first concerns the nature of knowledge and how we come to know things. The second source concerns the nature and representation of mental life. References: Driscoll, M. P. (2000). *Psychology of learning for instruction* (2nd Ed.). Needham Heights, MA: Allyn & Bacon.

Which of the following is true for the Student Version above?

- o **Word-for-Word plagiarism**
- o **Paraphrasing plagiarism**
- o **This is not plagiarism**

Item 4

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version (written in 2002)**
The technological tools available today for creating computer-based learning materials are incredibly more powerful than those introduced just a few years ago. We can make our own movies with camcorders in our homes; we can publish our own books. Soon teachers and students will be able to use computer-video technology to produce their own learning materials. All it takes is time, know-how, and some funds. References: Frick, T. (1991). *Restructuring education through technology.* Bloomington, IN: Phi Delta Kappa Educational Foundation.	Computers are so powerful that K-12 educators and students are now able to produce their own multimedia and Web-based learning materials. They just need to take the time required to learn to use the authoring tools and related technologies such as digital cameras and camcorders. References: Frick, T. (1991). *Restructuring education through technology.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Which of the following is true for the Student Version above?

- o **Word-for-Word plagiarism**
- o **Paraphrasing plagiarism**
- o **This is not plagiarism**

Item 5

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version**
The philosophical position known as *constructivism* views knowledge as a human construction. The various perspectives within *constructivism* are based on the premise that knowledge is not part of an objective, external reality that is separate from the individual. Instead, human knowledge, whether the bodies of content in public disciplines (such as mathematics or sociology) or knowledge of the individual learner; is a human construction. References: Gredler, M. E. (2001). *Learning and instruction: Theory into practice* (4th Ed.). Upper Saddle River, NJ: Prentice-Hall.	Does knowledge exist outside of, or separate from, the individual who knows? Constructivists hold that human knowledge, whether the bodies of content in public disciplines (such as mathematics or sociology) or knowledge of the individual learner; is a human construction (Gredler, 2001). References: Gredler, M. E. (2001). *Learning and instruction: Theory into practice* (4th Ed.). Upper Saddle River, NJ: Prentice-Hall.

Which of the following is true for the Student Version above?

- **Word-for-Word plagiarism**
- **Paraphrasing plagiarism**
- **This is not plagiarism**

Item 6

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version**
Major changes within organizations are usually initiated by those who are in power. Such decision-makers sponsor the change and then appoint someone else - perhaps the director of training - to be responsible for implementing and managing change. Whether the appointed change agent is in training development or not, there is often the implicit assumption that training will "solve the problem." And, indeed, training may solve part of the problem.... The result is that potentially effective innovations suffer misuse, or even no use, in the hands of uncommitted users. References: Dormant, D. (1986). The ABCDs of managing change. *In Introduction to Performance Technology* (p. 238-256). Washington, D.C.: National Society of Performance and Instruction.	When major changes are initiated in organizations, "... there is often the implicit assumption that training will 'solve the problem.' And, indeed, training may solve part of the problem." (Dormant, 1986, p. 238). References: Dormant, D. (1986). The ABCDs of managing change. In *Introduction to Performance Technology* (p. 238-256). Washington, D.C.: National Society of Performance and Instruction.

Which of the following is true for the Student Version above?

- **Word-for-Word plagiarism**
- **Paraphrasing plagiarism**
- **This is not plagiarism**

Item 7

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version**

The philosophical position known as *constructivism* views knowledge as a human construction. The various perspectives within constructivism are based on the premise that knowledge is not part of an objective, external reality that is separate from the individual. Instead, human knowledge, whether the bodies of content in public disciplines (such as mathematics or sociology) or knowledge of the individual learner; is a human construction. References: Gredler, M. E. (2001). *Learning and instruction: Theory into practice* (4th Ed.). Upper Saddle, NJ: Prentice-Hall.	The philosophical position known as *constructivism* views knowledge as a human construction. The various perspectives within constructivism are based on the premise that knowledge is not part of an objective, external reality that is separate from the individual. Instead, human knowledge is a human construction.

Which of the following is true for the Student Version above?

- o **Word-for-Word plagiarism**
- o **Paraphrasing plagiarism**
- o **This is not plagiarism**

Item 8

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version**
There is a desperate need for theorists and researchers to generate and refine a new breed of learning-focused instructional design theories that help educators and trainers to meet those needs, (i.e., that focus on learning and that foster development of initiative, teamwork, thinking skills, and diversity). The health of instructional-design theory also depends on its ability to involve stakeholders in the design process. References: Reigeluth, C. M. (1999). What is instructional design theory and how is it changing? In C. M. Reigeluth (Ed.), *Instructional-design theories and models volume II: A new paradigm of instructional theory.* Mahwah, NJ: Lawrence Erlbaum Associates.	We need theorists and researchers to generate and refine learning-focused instructional design theories. Such theories will help educators and trainers to meet needs that focus on learning and that foster development of initiative, teamwork, thinking skills, and diversity. Instructional-design theory must involve stakeholders in the design process. References: Reigeluth, C. M. (1999). What is instructional design theory and how is it changing? In C. M. Reigeluth (Ed.), *Instructional-design theories and models volume II: A new paradigm of instructional theory.* Mahwah, NJ: Lawrence Erlbaum Associates.

Which of the following is true for the Student Version above?

- o **Word-for-Word plagiarism**
- o **Paraphrasing plagiarism**
- o **This is not plagiarism**

Item 9

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version**
Instructional designers typically employ models to guide their day-to-day work. Due to the increased practice of the systematic design of instruction in a growing number of settings, available models become more and more proliferated, focusing on particular types and contexts of learning, particular groups of learners or designers, or particular instructional units (either whole curricula or individual modules or lessons.) The main goal of any instructional design process is to construct a learning environment in order to provide learners with the conditions that support desired learning processes. References: Merriënboer, J. J. van. (1997). *Training complex cognitive skills.* Englewood Cliffs, NJ: Educational Technology Publications.	"The main goal of any instructional design process is to construct a learning environment in order to provide learners with the conditions that support desired learning processes" (van Merriënboer, 1997, p. 2). Process models proliferate because more and more designers generate models that focus on specific contexts, learners, or even units of instruction, according to van Merriënboer. References: Merriënboer, J. J. van. (1997). *Training complex cognitive skills.* Englewood Cliffs, NJ: Educational Technology Publications.

Which of the following is true for the Student Version above?

- **Word-for-Word plagiarism**
- **Paraphrasing plagiarism**
- **This is not plagiarism**

Item 10

In the case below, the original source material is given along with a sample of student work. Determine the type of plagiarism by clicking the appropriate radio button.

Original Source Material	**Student Version**
Learning is a complex set of processes that may vary according to the developmental level of the learner, the nature of the task, and the context in which the learning is to occur. As already indicated, no one theory can capture all the variables involved in learning. References:	A learning theory is made up of a set of constructs linking observed changes in performance with whatever is thought to bring about those changes. Therefore since learning is a complex set of processes that may vary according to the developmental level of the learner, the nature of the task, and the context in which the learning is to occur, it is apparent that no one theory can capture all the variables involved in

Gredler, M. E. (2001). *Learning and instruction: Theory into practice* (4th Ed.). Upper Saddle, NJ: Prentice-Hall. --- A learning theory, there, comprises a set of constructs linking observed changes in performance with what is thought to bring about those changes. References: Driscoll, M. P. (2000). *Psychology of learning for instruction* (2nd Ed.). Needham Heights, MA: Allyn & Bacon.	learning.

Which of the following is true for the Student Version above?

- **Word-for-Word plagiarism**
- **Paraphrasing plagiarism**
- **This is not plagiarism**

Evaluate my answers for a possible certificate

CURRICULUM VITA

Andrew Barrett
www.andrewbarrett.ca

EDUCATION

2015 **PhD.**
School of Education, Indiana University
Dissertation: *Facilitating Variable-Length Computerized Classification Testing Via Automatic Racing Calibration Heuristics*
Supervisor: Dr. Theodore W. Frick
Major: Instructional Systems Technology
Minor: Nonprofit Management

2005 **Master of Science in Education, Curriculum and Instruction: Instructional Technology**
Southern Illinois University

1999 **Bachelor of Science, Computing and Information Science**
Queen's University

PROFESSIONAL EXPERIENCE

2013 – present **Assistant Director, Educational Development Centre,** Carleton University, Ottawa, Ontario

- Led campus wide Educational Technology, Instructional Design, and Educational Development teams
- Initiated, directed, and contributed to initiatives aimed at promoting teaching excellence among Faculty, Contract Instructors, and TAs
- Led initiatives to create open educational resources to help educators shift to teaching in blended and online contexts
- Oversaw design and development of new Carleton online courses
- Played leadership role in leveraging campus educational technology infrastructure to develop and deploy online training to university staff

2012 – 2013 **Manager, Instructional Innovation,** Carleton University, Ottawa, Ontario

- Oversaw Educational Technology team responsible university's successful migration to new learning management system
- Facilitated workshops with hundreds of Carleton educators to promote the effective use of technology generally and the learning management system specifically

2010 – 2011 **Graduate Assistant,** Office of Education Technology Services, School of Education, Indiana University, Bloomington, Indiana

- Created web-based systems to support performance evaluations, program evaluations, and the sharing of institutional knowledge

2007 Summer **Instructional Designer,** Centre for Healthy Development through Sport and Physical Activity, Brock University, St. Catharines, Ontario

- Created digital instructional materials for an ongoing project to promote life skills among youth in Central America

2006 Summer **Instructional Designer,** Option Six Corporation, Bloomington, Indiana

- Led the instructional design of two e-learning courses for Microsoft software developers

2005 Summer **Instructional Designer,** SchoolCenter, Carbondale, Illinois

- Designed and developed classroom training, manuals, and job aids for K-12 learning management system product

2001 – 2003 **Senior Associate of Technology,** Sapient Corporation, Toronto, Ontario

- Played key project leadership and IT roles on large-scale, web projects for clients including Union Gas and Scotiabank

1999 – 2001 **Associate of Technology,** Sapient Corporation, Chicago, Illinois

- Developed web-based software and led quality assurance efforts for clients including Dow Jones and IBM

1998 Summer **Software Developer,** Manta Corporation (now Halogen), Ottawa, Ontario

- Designed and implemented a bug-tracking system for web-based survey product

1996 – 1997 **Software Developer,** Information Technology Services, Queen's University, Kingston, Ontario

- Designed and implemented online course discussion tools used in several undergraduate programs

TEACHING EXPERIENCE

2012 – present **Workshop Facilitator,** Educational Development Centre, Carleton University, Ottawa, Ontario

- Facilitated dozens of workshops on topics including the integration of educational technology, effective assessment practice, course design, and increasing student engagement in various contexts

2012 Fall — **Volunteer,** Counseling and Educational Psychology Department, School of Education, Indiana University, Bloomington, Indiana
- Supported the creation of online instructional videos for online course EDUC P507: *Assessment in Schools*
- Helped to facilitate initial online discussions via course Wiki tool

2009 Fall — **Seminar Instructor,** Office of the Vice Provost for Undergraduate Education, Indiana University, Bloomington, Indiana
- Designed and taught COLL X111 A: *Teaching Young Children* and COLL X111 B: *Get Your Game On*

2005 – 2009 — **Assistant Instructor**, Department of Instructional Systems Technology, School of Education, Indiana University, Bloomington, Indiana
- Taught and helped redesign EDUC W200: *Using Computers in Education*
- Designed and taught EDUC R311: *Introduction to Instructional Systems Technology*
- Taught and helped redesign EDUC W201: *Beginning Educational Technology Skills*
- Managed technology and computing lab available to pre-service teachers

2003 – 2005 — **Graduate Assistant**, Department of Instructional Technology and Design, School of Curriculum and Instruction, Southern Illinois University, Carbondale, Illinois
- Taught and helped redesign EDUC 315: *Teaching with Technology*
- Supported faculty research, publication, presentation, and software development efforts
- Managed technology and computing lab available to pre-service teachers

PUBLICATIONS, PEER REVIEWED

Howard, C.D., Barrett, A.F. & Frick, T.W. (2010) Anonymity to promote peer feedback: Pre-service teachers' comments in asynchronous computer-mediated communication. *Journal of Educational Computing Research,* 43(1), 89-112.

PRESENTATIONS, PAPERS, AND POSTERS

Barrett, A.F. & Frick, T.W. (2014). *Featured Research - Facilitating Variable-Length Computerized Classification Testing In Massively Open Online Contexts Via Automatic Racing Calibration Heuristics*. Paper presented at the annual conference of the Association for Educational Communications & Technology, Jacksonville, Florida.

Frick, T.W. & Barrett, A.F. (2014). *Defeating the Cheating: Redesigning a Web-based Mastery Test during Heavy Usage.* Paper presented at the annual conference of the Association for Educational Communications & Technology, Jacksonville, Florida.

Barrett, A.F., Polovina-Vukovic, D., Davies, K. W., Sabra, S. (2014). *Perspectives on Blended Approaches to Faculty Development*. Presented at the annual conference of the Society for Teaching and Learning in Higher Education, Kingston, Ontario.

Barrett, A.F., Mackie, K. (2014). *Blended and Online Teaching Certificate and Open Educational Resources*. Presented at the annual conference Advancing Learning, Barrie, Ontario.

Barrett, A.F., Mackie, K. (2014). *Creating a Blended Teaching Certificate for the Postsecondary Teaching & Learning Community*. Presented at the annual conference CONNECT, Niagara Falls, Ontario.

Sabra, S. & Barrett, A.F. (2013). *Sustaining Learning, Integrity, and Community: Leveraging Course Design for Life-long Commitments to Social Responsibility and Ethical Conduct*. Workshop presented at the annual pre-conference of the Society for Teaching and Learning in Higher Education, Cape Breton, Nova Scotia.

Barrett, A.F. (2012). *Web 2.0 Based Participatory Assessment of TPACK in EdTech Classrooms*. Paper presented at the annual conference of the Canadian Society for the Study of Education, Waterloo, Ontario.

Barrett, A.F. (2012). *A Classroom Assessment Friendly Approach to Computer Adaptive Testing*. Paper presented at the annual conference of the Canadian Society for the Study of Education, Waterloo, Ontario.

Barrett, A.F. & Frick, T.W. (2011). *A Design Case: Improving an Online Plagiarism Tutorial by Preventing Cheating and Supporting Mobile Access*. Paper presented at the annual conference of the Association for Educational Communications & Technology, Jacksonville, Florida.

Barrett, A.F. & Frick, T.W. (2011). *Bridging the Divide Between Assessment Practice in Low and High-Stakes Contexts*. Paper presented at the annual conference of the Association of Educational Communications and Technology, Jacksonville, Florida.

Frick, T.W., Barrett, A.F., Gray, C.M., Howard, C.D. & Jung, J. (2011). *Presidential Panel: Preparing Students in Instructional Design and Technology to Become Skilled Researchers: Multiple Approaches*. Presented at the annual conference of the Association of Educational Communications and Technology, Jacksonville, Florida.

Frick, T.W., Myers, R. D., Howard, C.D., & Barrett, A.F. (2011) *Applications of MAPSAT in Educational Research: Map & Analyze Patterns & Structures Across Time*. Paper presented at the annual conference of the Association for Educational Communications & Technology, Jacksonville, Florida.

Barrett, A.F., Howard, C.D., & Frick, T.W. (2011). *Product Quality and Conditions of Anonymity In Formative Peer Assessment: Preservice Teachers' Asynchronous Computer-Mediated Communication*. Paper presented at the annual conference of the American Educational Research Association, New Orleans, Louisiana.

Barrett, A.F., Frick, T.W. (2011). *Facilitating Variable-Length Computerized Classification Testing in Instructional Contexts via Automatic Racing Calibration Heuristics*. Poster presented at the annual conference of the International Association of Computer Adaptive Testing, Pacific Grove, California.

Barrett, A.F. (2010). *Calibration Sample Size and Incorrect Identification of Problematic Test Items in Criterion-Referenced Testing*. Paper presented at the annual conference of the Association for Educational Communications & Technology, Anaheim, California.

Barrett, A.F. (2010). *Patterns of Technological Pedagogical Knowledge and Self-Efficacy in Preservice Teachers*. Paper presented at the annual conference of the Association for Educational Communications & Technology, Anaheim, California.

Barrett, A.F. (2010). *Measuring Teachers' Technological Pedagogical Knowledge*. Paper presented at the annual conference of the Canadian Society for the Study of Education, Montreal, Québec.

Barrett, A.F., Howard, C.D. & Frick, T.W. (2010). *Product Quality In Formative Peer Assessment: How Preservice Teachers' Computer-Mediated Feedback Changes With Peer Product Quality*. Paper presented at the annual conference of the Canadian Society for the Study of Education, Montreal, Québec.

Barrett, A.F. (2010). *Measuring TPK component of TPACK: An Alternative to Self-Assessment*. Paper presented at the annual conference of the American Educational Research Association, Denver, Colorado.

Barrett, A.F., & Howard, C.D. (2010). *Validity of Computer Mediated Formative Peer Assessment: Pre-service Teacher's Comments in Asynchronous CMC*. Roundtable presentation at the annual conference of the American Educational Research Association, Denver, Colorado.
Barrett, A.F., & Howard, C.D. (2009). *Product Quality in Online Peer Assessment: Pre-service Teacher's Comments in Asynchronous CMC*. Paper presented at the annual conference of the Association for Educational Communications & Technology, Louisville, Kentucky.

Barrett, A.F. (2009). *Development of an Assessment to Measure Teachers Educational Technology Knowledge*. Roundtable presented at the Association for Educational Communications & Technology Conference, Louisville, Kentucky.

Frick, T.W., Howard, C.D., Barrett, A.F., Enfield, J.W., & Myers, R. D. (2009). *Alternative Research Methods: MAPSAT Your Data to Prevent Aggregation Aggravation*. Paper presented at the annual conference of the Association for Educational Communications & Technology, Louisville, Kentucky.

Howard, C.D. & Barrett, A.F. (2009). *Anonymous Discussions: How learners commented on each others' work in a protected space*. Paper presented at the annual conference of the Association for Educational Communications & Technology, Louisville, Kentucky.

Howard, C.D., & Barrett, A.F. (2009). *Anonymity in Online Peer Reviews: Preservice Teachers' Online Comments in Two Critique Designs*. Paper presented at the annual conference of the American Educational Research Association, San Diego, California.

Barrett, A.F. (2008). *Leveraging ICT to Promote Multidisciplinary Environmentally Focused Authentic Educational Projects*. Poster presented at the annual conference of the North American Association for Environmental Education. Wichita, Kansas.

Barrett, A.F. (2008). *The Problems (And Opportunities) Related to Assessing Environmental Literacy*. Poster presented at the annual conference of the North American Association for Environmental Education. Wichita, Kansas.

SERVICE

Department

2006 – 2011	IST Student Orientation Week Volunteer
2006 – 2009	Indiana University IST Department Conference Volunteer
2006 – 2007	Indiana University IST Department Graduate Student Association President
2003 – 2004	Southern Illinois University Graduate Student Association President

School

2011 – present	Online survey development services to the Indiana University Student Advocates Office as part of a service evaluation project
2007	Balanced Scorecard Proposal for the Indiana University Board of Trustees created for SPEA V602 as part of a group project
2007	Workshop to help pre-service teachers at Indiana University's School of Education with the creation of their electronic portfolios
2004	Creation of proposed brochure for new program in School of Education at SIU as part of a course project

Professional

2012	STLHE Conference Proposal Review
2010 – 2012	CSSE Conference Proposal Reviewer and Session Facilitator
2010 – 2011	AECT Conference Session Facilitator
2010	AERA Conference Session Facilitator
2010	Treasure Nominee for Systemic Change Division in AECT
2004	ISPI Conference Volunteer

Community

2013 – present	Webmaster of gloucesterlions.ca for the Gloucester Lions Club, a service club in Ottawa, Ontario that does community work and provides students with service learning opportunities.

2011 – present	Co-founder of unlocklearning.org, an educational organization whose mission is to address equity in both formal and informal education by researching and developing high-quality, open, and mobile educational resources
2006 – present	Webmaster of niagarapenguins.org for Brock Niagara Penguins, an athletic team for youth with a disability in St. Catharines, Ontario that also provides service learning opportunities for Brock students

PROFESSIONAL MEMBERSHIPS

2012 – present	Society for Teaching and Learning in Higher Education
2009 – present	Canadian Society for the Study of Education
2008 – present	American Educational Research Association
2008 – present	Association for Educational Communications & Technology
2007 – 2008	North American Association for Environmental Education

AWARDS

2012	Team Service Excellence Award, Learning Management System Migration Team, Carleton University, Ottawa, Ontario
2008	IST/AVC Summer Fellowship, Instructional Systems Technology, School of Education, Indiana University, Bloomington, Indiana
2001	Core Value Award - Relationships, Sapient Corporation, Toronto, Ontario

INFORMATION TECHNOLOGY KNOWLEDGE

Programming	HTML, CSS, JavaScript, Python, PHP, Java, MySQL, and more
Applications	SPSS, Adobe Creative Suite, MS Office, Google Docs, and more
Web Systems	Moodle, WordPress, Mahara, Sakai, Badges, Blackboard, and more

LANGUAGES Native speaker of English. Earned French Bilingual Certificate through immersion programs at Emily Carr Middle School and Gloucester High School in Ottawa, Ontario.