BENCHMARK EXAM UTILIZATION IN CALIFORNIA PUBLIC MIDDLE SCHOOLS AND STANDARDIZED TEST SCORES: A NON-EXPERIMENTAL CORRELATIONAL STUDY

by

Michael F. Marcos

Liberty University

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Education

Liberty University

2015

BENCHMARK EXAM UTILIZATION IN CALIFORNIA PUBLIC MIDDLE SCHOOLS AND STANDARDIZED TEST SCORES: A NON-EXPERIMENTAL CORRELATIONAL STUDY

by Michael F. Marcos

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Education

Liberty University, Lynchburg, VA

2015

APPROVED BY:

Jared Bigham, Ed.D., Committee Chair

Justin Walker, Ph.D., Committee Member

Jason Bell, Ed.D., Committee Member

Scott Watson, Ph.D., Associate Dean, Advanced Programs

ABSTRACT

This quantitative non-experimental correlational research study intends to determine whether there was a correlation between the utilization of benchmark exams and state standardized test performance. Strictly focusing on schools that utilize benchmark exams, for how long they've utilized the exams, how frequently the exams are administered, how the exams are created, whether or not their administration is mandatory, and teacher satisfaction with the benchmark exams were considered as predictor variables. The Academic Performance Index (API) score change between the 2011 base API score and the 2012 growth API score was the criterion variable. The sample consisted of ninety-three California public middle schools. The sample was a balanced representation of academically low, mid, and high-performing schools as measured by the Academic Performance Index (API). The use of a multiple linear regression analysis quantified whether there is a correlation between how benchmark exams are utilized and Academic Performance Index (API) scores as measured by year-over-year value-added growth. The result of the study failed to indicate a significant positive linear relationship amongst the variables. As such, for future research, it is recommended to greatly expand the sample size by adding California public elementary and high schools. The goal would be to include enough participating schools that the percentage of schools not utilizing benchmark exams at each grade level would be comparable to those that do.

Keywords: Benchmark Exams, Common Assessments, Middle Schools, Academic Performance Index (API), Adequate Yearly Progress (AYP), California Standards Test (CST), Program Improvement (PI), Standardized Testing and Reporting (STAR).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Academic Performance Index (API)

Adequate Yearly Progress (AYP)

California Standards Test (CST)

English language arts (ELA)

Local Education Agency (LEA)

Massachusetts Comprehensive Assessment System (MCAS)

No Child Left Behind Act of 2001 (NCLB)

Program Improvement (PI)

Standardized Testing and Reporting (STAR)

CHAPTER ONE: INTRODUCTION

Background

Since the passage of No Child Left Behind in 2001, high stakes standardized tests have become a mandatory accountability tool in all public schools in the United States. States are required to rate schools based on test results in order to receive federal funds (Kastenbaum, 2012). In California, the introduction of the Standardized Testing and Reporting (STAR) program mandated standardized testing in 1998 (Ed-data, 2014). Students were given these exams to display their level of content proficiency and assess how well their teachers and schools prepared them. Ultimately, it is the school that faces severe consequences if their students fail to perform at a level deemed appropriate by the state and federal government. More disconcerting, according to those opposed to the emphasis placed on standardized tests, is the notion that test scores can be used as an accurate indicator of teacher efficacy. This belief dramatically increases the pressure on teachers to produce desirable, quantified outcomes, as their job may very well depend on it. "I find it the most absurd thing in the world. I don't know anyone who thinks they're valid" said Principal Anna Allanbrook at Public School 146 in Brooklyn, New York. Kastenbaum (2012) stated, "So the morale is down because teachers are worried that people who don't really know their work will make decisions about their jobs" (para. 4). Additionally, those same critics assess that, when such high stakes are attached to standardized tests, the tests become corrupted; Kastenbaum (2012) also stated, "People cheat. We've seen major cheating scandals in both Atlanta, Georgia, and Washington, D.C, but in many other districts as well" (para. 13).

Each year, the federal government sets targets determining:

the minimum percentages of students who are required to meet or exceed the proficient level on the statewide assessments used for AYP (Adequate Yearly Progress). These targets apply to all schools of the same type (elementary, middle, or high school), to all LEAs (Local Education Agency) of the same type (elementary, unified, or high school district), to the state, and to all numerically significant groups. (California Department of Education, 2011)

Schools, nationwide, that fail to meet proficiency targets, as measured by Adequate Yearly Progress (AYP), are identified as program improvement and placed in corrective action. In California, specifically, schools are also assigned an Academic Performance Index (API) score once student performance on the California Standards Tests (CSTs) is calculated. This score ranges from 200-1000, with 1000 being the best possible score and an indication that all students tested have reached a level of advanced content mastery (California Department of Education, 2011). This API score is generally viewed as the school's report card grade. It is a heavily weighted factor in the perception of the school's prestige, or lack thereof, in the eyes of community stakeholders and universities. Every year, prior to the administration of the CSTs, the California Department of Education sets an API growth target for each school. The ultimate goal was for all schools in California to have achieved an API of 800 or higher by the year 2014. This was an unrealistic goal to begin with; that deadline has now passed and schools and school districts that have failed to achieve that goal are losing autonomous control over daily operations. Therefore, more and more schools are looking for ways to ensure student success on these end-of-year exams in order to exit program improvement and regain their operational and decision-making independence.

One approach some schools use is to create a series of benchmark exams, which are common assessments that are administered throughout the course of the year leading up to the

end-of-year standardized tests. In California, unlike many states in America, these exams are not mandatory. Schools that utilize these benchmark exams would argue that they lend themselves to increased student achievement on standardized tests resulting from improved teacher pacing, revising of instructional practices, more goal-oriented learning, and better standards alignment due to the ability of these exams to generate immediate feedback. Advocates assert that, "the primary purpose of benchmark assessments is to inform teaching and learning" (Hefflin, 2009, p. 11). Critics cite their use as one more barrage of standardized tests in an era where teacher creativity has been hampered as a result of needing to teach to the test. These same critics call for the elimination or sharp reduction of this form of testing. Many argue that benchmark exams are not useful for improving instruction and instead are ineffective low-quality tests (Nelson, 2013).

Problem Statement

In today's educational environment, schools continuously face a barrage of challenges related to budget crises and a rapidly evolving student clientele with unique needs. Regardless of these barriers, schools are being held to a higher level of federal and state accountability than ever before, and a uniformly applied one at that. The overwhelming majority is based on results of high stakes standardized tests. Under the No Child Left Behind Act, test scores impact how much funding a school gets from the government, as well as how much autonomy a school has (Evans, 2013). District and site administrations are looking for ways to boost student performance on standardized tests in order to avoid corrective actions levied against them by the state and federal government. That said, not all government officials are in favor of the weight carried by these exams. In an apparent rebuke of the No Child Left Behind Act of 2011, current US Secretary of Education, Arne Duncan, said that much of the criticism of standardized testing

is warranted. Duncan stated, “State assessments in mathematics and English often fail to capture the full spectrum of what students know and can do…Students, parents, and educators know there is much more to a sound education than picking the right answer on a multiple-choice question” (Evans, 2013).

Although some schools in California have implemented the use of benchmark exams, the State Department of Education has not mandated it. Instead, California has opted for localized decision-making regarding whether or not to utilize benchmark exams. Opponents of this type of systemic assessment, such as former Texas State Senator Ted Lyon, argue that the risk is not worth the reward, citing the significant loss of instructional time. For example, high school students in Texas spend between 29 and 45 days a year taking standardized tests. In Tennessee, students spend six weeks in testing a year, and California’s students spend four, according to PolicyMic.com. These numbers do not include the weeks and months spent on test preparation classes and benchmark practice exams (Evans, 2013).

Previous studies have tried to identify whether or not benchmark exams can serve as an accurate predictor of end-of-year standardized test scores. In Massachusetts, benchmark exams are mandated by the state yet, despite the ubiquity of benchmark testing, research conducted by Mathematica Policy Research in Boston on these tests failed to demonstrate a positive and statistically significant impact on improving student achievement on end-of-year exams (Furgeson & Gill, 2012). The problem is there has yet to be a study conducted in California middle schools to determine whether or not benchmark exams improve standardized test scores and, if so, what logistical variables of utilization generate a significant positive correlation, if any. Therefore, the gap in the literature remains since there has still been no conclusive study identifying whether or not there is a direct correlation between how benchmark exams are

utilized and student performance on state standardized tests, specifically the California Standards Test.

Purpose Statement

The purpose of this quantitative, non-experimental, correlational research study was to determine whether or not there is a direct relationship between how benchmark exams are utilized and end-of-year standardized test scores, specifically on public middle school campuses in California. As Gall, Gall and Borg (2007) states, "The basic design in correlational research is very simple, involving nothing more than collecting data on two or more variables for each individual in a sample and computing a correlation coefficient" (p. 335). This is the appropriate selection for this study, as it will identify the effect the predictor variables of benchmark exam utilization have on standardized test scores, if it is determined there is one. The predictor variables of this study are how long the benchmark exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams. The criterion variable will be the score change between the 2011 base and 2012 growth API scores of the schools. Data analysis will identify whether or not a linear combination of the predictor variables impacts student performance on the CSTs as measured by the criterion variable.

The participating sample consisted of 93 public middle schools. All schools operate in the state of California and take the California Standards Tests after the same number of days from the first day of their school year. All schools utilized benchmark exams in some capacity. Participating schools will equally represent the categories of high-performing (API > 800), mid-performing (700-799), or low-performing (below 700) based on API scores. Within each category, subgroup demographics will be comparable. It is assumed that teachers and students

do not cheat on the standardized tests and put forth their best effort on benchmark exams and end-of-year standardized tests.

Significance of the Study

In addition to state-mandated end-of-year standardized tests, almost all districts have some type of benchmark exams designed to measure student progress toward passing the state-mandated test over the course of the school year (Heppen & Jones, 2011). But, unlike the state-mandated end-of-year exams, results from these assessments are available during the school year. Advocates of benchmark exams assert that it is precisely this data that are believed to help improve scores on state-mandated tests when used to target instructional interventions, to review and reteach, to regroup students, to identify students for tutoring, and to share results with students (Nelson, 2013).

Critics of benchmark exams argue there is no statistical evidence supporting their efficacy in achieving the ultimate goal of increased student performance on state-mandated end-of-year standardized tests. Because of this, these same critics rail against the financial and instructional time costs as being grossly wasteful and counterproductive. The American Federation of Teachers referenced a comprehensive study of two mid-size urban public school districts, one in the mid-west and one on the east coast. The purpose of this study was to quantify the cost of these districts' financial and instructional time expenditures on their benchmark exam and standardized testing programs. The results were staggering, particularly when considering the fact that these are tax-payer-funded school districts. In calculating the average of the two districts, it was determined that the annual testing programs cost $1000 per pupil per year and 40 minutes of lost instructional time each day (Nelson, 2013).

Should this study indicate there is a direct positive correlation between how benchmark exams are utilized and student performance on end-of-year standardized tests, schools and school districts would now have significant verification that there is a strategy proven to yield positive results. As the stakes rise and school districts, schools, administrators and teachers all find themselves under increasingly immense pressure, the results of this study could influence the educational landscape of California and, potentially, compel state government officials to make the use of benchmark exams mandatory as they are in other states. If no direct positive correlation between the utilization of benchmark exams and student performance on end-of-year standardized tests is statistically indicated, school districts will have valuable information on which to amend financial expenditures, professional development, teacher evaluation, and data collection protocols.

Research Question

This research question for this study is:

RQ1: Can changes in the API scores of California public middle schools that utilize benchmark exams be predicted from a linear combination of how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams?

Null Hypothesis

The null hypothesis for this study is:

$H_0$1: Changes in the API scores of California public middle schools that utilize benchmark exams cannot be predicted from a linear combination of how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams.

Definitions

1. *Academic Performance Index (API)* – An annual measure of test score performance of schools and districts (Ed-data, 2014).
2. *Adequate yearly progress (AYP)* – The measure by which schools, districts, and states are held accountable for student performance under Title I of the No Child Left Behind Act of 2001 (Education Week, 2011).
3. *Benchmark Exams* – Administered at several points throughout the instructional year, in order to give insight as to how students are preparing for the coming assessments (ASCD, 2005).
4. *Corrective Action* – Steps a district must take to ensure student achievement in the core academic subjects (Oceanside Unified School District, 2015).
5. *California Standards Tests (CSTs)* – Developed by California educators and test developers specifically to measure students' progress toward achieving California's state-adopted academic content standards in English–language arts (ELA), mathematics, science, and history–social science (Standardized Testing and Reporting Program, 2014).
6. *Local Education Agency (LEA)* - A public board of education or other public authority legally constituted within a state for either administrative control or direction of, or to perform a service function for, public elementary schools or secondary schools in a city, county, township, school district, or other political subdivision (US Department of Education, 2012).
7. *No Child Left Behind Act of 2001 (NCLB)* – Federal legislation that enacts the theories of standards-based education reform (US Legal, 2015).

8. *Program Improvement (PI)* – A school that fails to make AYP for two consecutive years is identified as such and must develop or revise an improvement plan, have the state approve the plan, and devote to professional development an amount equivalent to 10 percent of its annual Title I funds (North Dakota Department of Public Instruction, 2015).
9. *Standardized Testing and Reporting (STAR)* – Authorized in 1998, students in grades two through eleven must take exams each spring in math, reading, writing, science, and history. In California, these exams are the CSTs. The STAR Program assesses how well schools and students are performing (California Department of Education, 2011).

CHAPTER TWO: LITERATURE REVIEW

Introduction

There is no shortage of opinions when it comes to the value of benchmark exams, common assessments, or any form of standardized testing, government-mandated or otherwise. Educational stakeholders, ranging from classroom teachers to politicians, hotly debate their merit as formative or summative assessments, time and cost-efficacy, use as an instructional tool, and, most relevant to this study, their ability to prepare students for success on end-of-year state exams. Whereas the bulk of the literature tends to be subjective scholarly articles, there have been the occasional studies attempting to determine the extent to which benchmark exams can serve as predictors on standardized tests, such as Sherman's (2008) dissertation on the Texas state-mandated benchmark exams. Although also not based in California, Hefflin's (2009) dissertation did attempt to study the relationship between the Pennsylvania state-mandated 4Sight benchmark assessments and student's performance on the end-of-year standardized tests. Perhaps the most similar study to this one took place when a research team used a quasi-experimental design to examine the effectiveness of a Massachusetts pilot program in which selected schools used quarterly benchmark exams aligned with state curriculum standards for middle-school mathematics. The researchers measured student test scores at 22 pilot program schools and 44 comparison schools on the eighth-grade Massachusetts Comprehensive Assessment System (MCAS) math exam from 2001 to 2007. The study identified whether or not the Massachusetts pilot program schools showed greater gains in student achievement than schools not in the program. After two years of program implementation (2006 and 2007), no statistically significant difference in test scores could be found between schools participating in the benchmark assessment pilot program and the comparison schools. That finding might,

however, reflect limitations in the data rather than the ineffectiveness of benchmark assessments. First, data is lacking on what benchmark assessment practices comparison schools may be using, because the study examined the impact of a particular structured benchmarking program. More than 70% of districts are doing some type of benchmark assessment, so it is possible that at least some of the comparison schools implemented their own version of benchmarking (Henderson, Petrosino, Guckenburg, & Hamilton, 2007). The researchers cautioned against drawing conclusions about benchmark assessments based on the project, which could not control for all school variables including leadership, student motivation, teacher training, and how the schools use the benchmark data (Ed-Evidence, 2009).

While similar in nature, the aforementioned studies leave a gap in the literature. In an effort to close it, this research study focused on California public middle schools where, as with the entire state, benchmark assessments are strictly optional. Because of the lack of standardization and mandate regarding California benchmark program, there is a presence of several variables regarding benchmark implementation. The analysis of the role of some of these variables, such as for how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams, along with the statistical calculations of California-specific API scores, further differentiated this study from others and make it a useful tool for educators across the state.

Conceptual or Theoretical Framework

The study is based on multiple theoretical frameworks, the first of which is Keller's (1983) ARCS Model of Motivational Design. Keller (1983) stated that in order to promote and sustain motivation in the learning process, students must go through four steps: attention,

relevance, confidence, and satisfaction. When applied to benchmark exams and standardized tests, the ARCS model highlights a potential limitation in the study in which a barrage of testing without sufficient student motivation and understanding of its relevance can have counterproductive results. "Keller & Litchfield (2002) assert[ed] that even the most accurate content and related activities can be ineffective without the systematic incorporation of motivation to improve student motivation" (Akdemir & Colakoglu, 2010, p. 87).

If teachers have negative attitudes towards the benchmark exams, their students will adopt the same views that could compromise their performance on these tests. Teachers must carefully design their instructional strategies to make sure students are motivated and see the relevance in the lesson or assessment tool. Further evidence supports the validity of Keller's (1983) ARCS model of motivational design, as Akdemir and Colakoglu (2010) stated, "Instructional design incorporating instructional and motivational components are critical to achieve learning goals" (p. 86).

The second relevant theoretical framework and the basis for student assessment continuums is Bloom's cognitive taxonomy. According to Bloom, Englehart, Furst, Hill, and Krathwohl (1956), a focus on higher-level thinking skills helps to engage young minds and attach relevance to their learning (Bloom et al., 1956). The original Taking Center Stage aligns Bloom's cognitive taxonomy to the continuum of assessment (Taking Center Stage, 2001) and highlights the fact that many professional learning communities use Bloom's taxonomy when developing or selecting common assessments (Taking Center Stage – Act II, 2010).

A common criticism of standardized assessment is that it is devoid of relevance and opportunities for students to utilize higher-level thinking skills. Negatively stereotyped as a regurgitation of rote-memorized facts, these assessments would barely make it onto Bloom's

(1956) taxonomy. One might argue that standardized assessments provide students with opportunities to display knowledge, the lowest level of the cognitive pyramid, but standardized assessment opponents would adamantly claim that the higher levels of Bloom's cognitive taxonomy – comprehension, application, analysis, synthesis, and evaluation – are nowhere to be found and cannot be assessed by filling in bubbles (Bloom et al., 1956).

Feedback for learning is also salient in the literature on motivation and self-efficacy (Heritage, 2010) and, with respect to these processes, Vygotsky's (1978) theory of the zone of proximal development has particular relevance. Vygotsky viewed learning as a social process in which learners collaborate with more expert others, such as teachers or peers, to develop cognitive structures that are still in the course of maturing and which are unlikely to fully mature without interaction with others (Heritage, 2010).

It could be asserted that a standardized assessment program deprives students of meaningful feedback and opportunities to collaboratively display whether meaning has been made or not. As these types of exams typically consist of multiple choice and true/false questions, the extent of the feedback students receive is limited to a red mark. Proponents of benchmark exams make the case that the ongoing data generated by systematic assessments provide teachers with the necessary information to provide meaningful feedback and adjust instruction accordingly.

These theoretical frameworks are the foundation upon which this study is conducted, and pose some powerful questions. Is the role of motivation significant when studying the impact of an educational tool that requires a student putting forth their best effort to accurately assess its efficacy? Can higher-level thinking skills, while widely viewed as a positive if not necessary quality of any assessment, be emphasized and truly assessed on standardized tests, or does the

need for them to be mass-produced and scored prove the priority in their analysis? And, finally, can the concept of learning as a social process in which meaningful feedback is provided from expert others coexist in a standardized assessment program consisting of benchmark and end-of-year state exams?

Review of the Literature

In order to fully contextualize both the positive and negative opinions of benchmark exams and their potential to affect meaningful student achievement growth, it must first be understood exactly what a benchmark assessment is and the rationale behind its origin and utilization.

Although there are many types of formative and summative assessments, one of the most common examples is the benchmark exam (FairTest, 2009). The past ten years have witnessed an explosion in the use of benchmark exams by school districts across the country. A primary reason for this rapid growth is the assumption that benchmark assessments can inform and improve instructional practice and thereby contribute to increased student achievement. Testing companies, states, and districts have become invested in selling or creating benchmark assessments and data management systems designed to help teachers, principals, and district leaders make sense of student data, identify areas of strengths and weaknesses, identify instructional strategies for targeted students, and much more. In an educational environment of reduced government funding, increasing class sizes, and dwindling resources, districts are keeping their benchmark assessment programs even under increasing pressure to cut budgets (Goertz, Olah, & Riggan, 2009).

It's no secret why districts are turning to benchmark tests. The No Child Left Behind Act, signed into law by President Bush in January 2002, and states' own accountability systems

have created a high-stakes environment in which both districts and schools can face penalties for failing to meet performance targets (Olson, 2005). California, in adherence to these federal mandates, has created content and performance standards for English-language arts and mathematics. By creating the performance standards, California has defined what a student should know and at what level of proficiency. Through the adoption of these standards, the state has clearly affirmed what content students need to acquire at each grade level. With these standards in place, student achievement and mastery of these standards are measured with the California Standards Tests (CST), which are criterion referenced tests developed specifically for California. As part of the state's accountability system, performance on the CST also constitutes the largest component of the school's Academic Performance Index (English-Language Arts Benchmark Tests, 2004).

In contrast with other countries, tests in the United States are often used to determine the curriculum in which students can enroll and whether they are promoted or allowed to graduate; whether teachers are tenured, continued, or fired; and whether schools are rewarded or sanctioned, or even reconstituted or closed (Darling-Hammond & Adamson, 2013). It is this group of teachers who are the fiercest opponents of benchmark exams, particularly in those states where they are mandatory. These teachers fear the recent push to evaluate and possibly even to determine the pay of teachers as a result of benchmark exams and standardized test scores. A further complication at the secondary level is that, "many specialty classes at the high school level lack a common assessment that could be used as a benchmark" (Haug, 2010, p. 1). Many teachers only teach these specialized, often singly offered courses, thereby giving them no comparative basis for such merit-based pay. With scores used to determine so many decisions,

critics of high-stakes standardized benchmark assessments explain that the incentives for teachers teaching to the test have become increasingly intense (Zehr, 2006).

Because of pressures to teach to the test in high stakes accountability systems, the additional costs for interim and benchmark testing have become viewed as mandatory. However, in many cases they may not improve the quality of assessment or leverage higher quality instruction because they are focused on raising scores on current state tests, which measure mostly low-level skills (Darling-Hammond & Adamson, 2013). Currently, the average state-testing system in reading and mathematics costs $25 to $27 per pupil. However, the pressures of meeting accountability requirements have caused states and localities to add additional interim and benchmark tests, as well as increasing spending for data systems and test preparation. In combination, these expenditures now average more than $50 per pupil. California's state testing program is one of the least expensive in the country, costing about $17 per pupil for assessments in English language arts (ELA) and mathematics, and just under $20 per pupil for the entire state testing program, including science and social studies. But, at an average of $15 per pupil, the interim tests appear to cost nearly as much as the state tests themselves (Darling-Hammond & Adamson, 2013). In total, these investments amount to many billions of dollars of educational investment that may not be leveraging the kinds of instruction required to meet the Common Core standards and to master 21st century skills (Topol, Olson, & Hennon 2013, p. 12). Money, time, and energy invested in benchmark assessments could divert attention from the more potent lever of changing what teachers do in classrooms each day, such as the types of questions they ask students and how they comment on students' papers (Olson, 2005).

Whereas the state government has set the curricular agenda and mandated the use of a specific end-of-year standardized test, benchmark exam use in California is completely optional. It is not required by the state, and there is no standardized California version of a benchmark exam. Therefore, a benchmark exam used in California is a locally customized, usually district-wide assessment administered periodically throughout the school year at specified times during a curriculum sequence to evaluate students' knowledge and skills relative to an explicit set of longer-term learning goals. Whether teachers, consulting firms, or test banks create the benchmark exams is left up to the local educational agency. Whether the exams are considered optional or mandatory and how many times they are administered each year is also left up the school or district. Typically, the exams are designed to measure the achievement of standards and performance objectives that are generally aligned with those specified at the state level and on the end-of-year state exam. However, local standards may also be included. The fundamental purpose of benchmark assessment is to provide information that can be used to guide instruction. Benchmark tests measure student mastery of standards targeted for instruction but can also inform instruction in cases where standards have not been mastered. They support interventions by identifying specific skills that students need to acquire in order to master standards targeted for instruction. Data generated from benchmark exams is valuable information that teachers can use to revise instructional practices according to areas of strength and weakness. Benchmark assessments provide this information in a cyclical manner, typically consisting of teaching, assessment, and intervention implemented intermittently throughout the school year (Bergan & Burnham, 2009). Although there is no state benchmark mandate, variables such as the frequency of utilization and how they are created are ultimately left to the local district or school (Herman, Osmundson, & Dietel, 2010). A typical California unified

public school district may reference benchmarks on their website or in promotional literature as part of an overarching assessment program. For example, one school district's website includes:

Benchmark Exams are given three times a year at key instructional points and prior to the annual administration of the California Standards Tests. The Benchmark Exams provide teachers, site administrators, and school district administrators tools to correlate data to the CST and to plan for areas needing improvement. (Norwalk-La Mirada Unified School District, 2015).

Because of the lack of standardization, there are inconsistencies across districts in regards to benchmark testing policies. In some districts, they are graded but not part of students' grades. The test scores will then show up on their report card as a separate category just so parents know and the students know what the grade is, but it doesn't have any effect on their class grade. A lot of this is decided by the teacher. The assessment program is not implemented county or state-wide (Abrams, McMillan, & Wetzel, 2010).

Now we're putting individual items in the hands of teachers, saying, 'You construct the test; make it as long or as short as you want.' Do we think they have the understanding to know how much stock they can put in the generalizations they make from such exams?" (Olson, 2005)

What might be the biggest impediment to standardized implementation of benchmark exams is the anti-testing attitude of many state officials. States are currently left to decide whether or not to mandate any type of common assessment beyond the federally mandated end-of-year standardized tests. Therefore, state departments of education and governors must favorably view the value of such benchmark exams in order to make such a mandate. One might assume all government entities view testing as an appropriate accountability tool, but this is clearly not the case. "'I'm open to it, but it needs to be without additional testing time,' Arkansas

Gov. Mike Beebe, a Democrat, said in an interview. ‘There is so much teaching to the test already, and we don't need any more of it’" (McNeil, 2008, p. 1).

Recently, the federal government has hesitantly waded into this debate. In his 2009 State of the Union address to Congress, President Obama stated,

I am calling on our nation’s Governors and state education chiefs to develop standards and assessments that don’t simply measure whether students can fill in a bubble on a test, but whether they possess 21st century skills like problem-solving and critical thinking, entrepreneurship, and creativity. (Darling-Hammond & Adamson, 2013)

With hundreds of millions of dollars earmarked for educational reform per the President’s Race to the Top initiative, representatives of the U.S. Department of Education asked for input from a group of educational experts as to how they could improve the country’s assessment systems. The overwhelming majority of the responses echoed this sentiment: “Teachers must be much more involved in the development, use, and possibly even in the scoring of assessments” (Sawchuk, 2009, p. 16).

In the eyes of critics, there are many problems with the use of benchmark assessments, including the fact that the assessment methods that teachers use are not effective in promoting good learning; grading practices tend to emphasize competition rather than personal improvement, and assessment feedback often has a negative impact, particularly on low-achieving students who are led to believe that they lack ability and are not able to learn (Black, Harrison, Marshall, & William, 2004). Students given feedback as marks are likely to see it as a way to compare themselves with others (ego involvement); those given only comments see it as helping them to improve (task involvement). Research shows that the latter group generally outperforms the former (Black et al., 2004). Another related point of contention is the issue of

cultural bias of standardized tests. With the population of English language learners rapidly increasing, particularly in states like California, schools and districts are now dealing with the reality that these students' scores can no longer be excluded from the data.

In the past, the answer to the standardized testing dilemma was to simply exclude students from testing who had been in the USA for a limited number of years or who were deemed not to have the English proficiency needed to participate in the testing process. (Butler & Stevens, 2001, p. 411).

Schools are now held accountable for all students' content proficiency, regardless of subgroup or demographic, lending an additional anti-testing voice to the mix.

The controversy over the use of benchmark exams and whether they are an attribute or detriment to student success continues to be a topic of much debate. Opponents frequently assail the unequal benefit across demographics. Designed to hold teachers and students accountable for content coverage, benchmark exams are intended to ensure that students know what they need to know and by when they need to know it. Developed to be a tool which levels the playing field,

> The problem is that such tests, ostensibly developed to 'leave no student behind', are in fact causing major segments of our student population to be left behind because the tests cause many to give up in hopelessness – just the opposite effect from that which politicians intended. (Stiggins, 2002, p. 759)

A troubling reinforcement of this argument is the fact that the use of benchmark assessments is disproportionately more prevalent in urban districts than in rural or suburban districts. In fact, 94% of research directors of urban school districts indicated that benchmark exams are administered in English language arts and mathematics (Topol et al., 2013). One urban school

district teacher opponent stated,

> I want the state to abandon its effort to turn me into a delivery system of approved information. I want it to support me and other teachers as we collaborate to create curriculum that deals forthrightly with social problems, that fights racism and social injustice. I want it to acknowledge the legitimacy of a multicultural curriculum of critical questions, complexity, multiple perspectives, and social imagination. I want it to admit that wisdom is more than information – that the world can't be chopped up into multiple-choice questions and that you can't bubble in the truth with a number-two pencil. (Bigelow, 1999, p. 40)

In a passionate criticism of current educational practices, a coalition of urban school district teachers claimed in a rally, "Standardization and centralization of curriculum testing is an effort to put an end to a cacophony of voices on what constitutes truth, knowledge and learning and what the young should be taught. It insists upon one set of answers" (Bigelow, 1999, p. 37).

Groups who have already and will certainly continue to profit from the push for localized benchmark exams are the test publishers and educational consulting firms. "While many schools have enacted teacher made in-house assessments and item banks, most enjoyed the convenience of commercially made benchmark tests" (Olson, 2005). "The market for formative assessments is now one of the fastest growing sectors of test publishing companies" (Casting, 2008, p. 3). A plethora of benchmark assessments is currently available to educators ranging from glossy, high-tech versions designed by testing companies that feature built-in data analysis tools and reports to locally developed, instructionally-driven assessments (Herman, Osmundson, & Dietel, 2010).

The quality of assessments has become a significant issue. Most states that were pursuing open-ended performance assessments in the 1990s dropped them during the NCLB era,

largely due to cost issues. Many states now rely solely on multiple-choice items on "bubble in" types of tests. Some are tests of rote memorization rather than measuring 21st century skills. Additionally, benchmark assessments and other types of systematic testing have been implemented. Little evidence exists that any of this increase in testing and test preparation is leading to improved student learning or critical thinking skills or the abilities that will prepare them for college and careers (Topol et al., 2013). "Teachers worry that to prepare our students for the tests, we will have to turn our classrooms into vast wading pools of information for students to memorize" (Bigelow, 1999, p. 37). "Because there is no way to predict precisely which facts will be sought on the state tests, teachers feel pressured to turn courses into a memory Olympics" (Bigelow, 1999, p. 38). These are not the 21st century skills the president called upon educators to develop.

Still, many teachers firmly stand by the usefulness of benchmark exams. Compared to end-of-year state exams, teachers found them to be more useful as they identified and corrected gaps in their teaching (Abrams et al., 2010). Proponents of benchmark assessments claim that when used correctly, these tests have the potential to give specific feedback on the academic areas in which individual students need the most assistance (Coffey, 2009). As one teacher stated,

It makes a difference in my instruction. I mean, I think I'm able to help students more that are having difficulty based on it. I am able to hone in on exactly where the problem is. I don't have to fish around....If I see a large number of my students missing in an area, I am going to try to re-teach it to the whole class using a different method. If it is only a couple of students, I will pull them aside and instruct one-on-one. (Abrams et al., 2010)

Teachers' ongoing use of assessment to guide and inform instruction – classroom formative assessment – can lead to statistically significant gains in student learning (Herman et al., 2010). Districts, schools, and teachers can use benchmark data to predict whether students, classes, schools, and districts are on course to meet specific year-end goals, such as if they will be classified as proficient on the end-of-year state test (Herman et al., 2010). Additionally, schools and districts may use benchmark results to allocate resources such as. time, staff, professional development, technical assistance, and special interventions (Herman et al., 2010). This is very valuable information as school districts continue to try to maximize the impact of continuously decreasing budgets during this era in which doing more with less is the onus placed upon educational leaders.

School districts worried about how students will perform on end-of-the-year state tests are increasingly administering benchmark assessments throughout the year to measure students' progress and provide teachers with data about how to adjust instruction. Teachers want their students to perform well on high stakes end-of-year assessments and thus tend to focus classroom curriculum and instruction on what will be assessed and to mimic assessment formats (Herman, 2009). Such tests typically are aligned to state or district standards for academic content and given anywhere from three to five times during the year, in some cases as often as monthly. Benchmarks, therefore, can also serve as pacing guides for teachers and schools, providing information on whether students have learned the curriculum they've just been taught (Olson, 2005). But, while many assessment experts agree that the idea of frequent testing of students to monitor their learning and adjust instruction is sound, some also warn that districts should take a close look at what they're getting for their money and how they are using such exams (Olson, 2005). Despite the glitz and gee-whiz appeal of such products, information about

their effectiveness in improving student learning is generally hard to come by (Herman & Baker, 2005).

I think it has definitely made us change the way we teach because you are looking for how can I teach this the most effectively and the fastest… that is the truth. You have got to hurry up and get through the curriculum so that you can get to the next thing so they get everything before the test. I do feel like sometimes I don't teach things as well as I used to because of the time constraints… We are sacrificing learning time for testing time… we leave very little time for actually teaching. These kids are losing four weeks out of the year of instructional time. (Abrams et al., 2010)

"Relying on high-stakes test results for instructional guidance is like trying to get to the Empire State Building with a map of the United States" (Supovitz & Klein, 2003, p. 1). "Receiving test data in July is like driving a school bus looking out of the rearview mirror. I can see where my students have been but I cannot see where we are going" (Salpeter, 2004, p. 30).

Research has consistently shown that the use of benchmark tests is a strongly recommended method to gauge mastery throughout the school year, provide teachers with diagnostic and prescriptive information, and provide students with test-taking skills (Action Learning Systems, 2004). A major premise in the development and use of benchmarks is that there is a positive relationship between scores a student receives on both the benchmark tests and the California Standards Tests. One way of expressing this relationship, for example is, if a student scores high on the benchmark test, that student should also score high on the CST (Action Learning Systems, 2004). "Benchmark assessments often serve four interrelated but distinct purposes: (a) communicate expectations for learning, (b) plan curriculum and instruction, (c) monitor and evaluate instructional and/or program effectiveness, and (d) predict future

performance" (Herman et al., 2010). Supporters also suggest that when benchmarks are created in alignment with state standards, they enable teachers to more accurately gauge student performance against district standards (Olson, 2005). For benchmark assessments that serve predictive purposes, the relationship between benchmark results and end-of-year state assessments is of high interest. If the benchmark assessment scores are highly correlated with proficiency levels on the end-of-year test, benchmark scores can be used to identify students who are likely to achieve proficiency and those who are not (Herman et al., 2010).

Aligning benchmark tests with state standards does not mean creating formative tests that mimic the content and format of the annual state tests as specifically as possible. Although a strategy of strict test preparation may boost test scores in the short term, available evidence suggests that early gains achieved in this way are not sustained in the long run (Herman & Baker, 2005). When a relationship has been established between performance on one or more benchmark assessments and performance on a statewide test, benchmark results can be used to assess the level of risk that a given student will not meet state standards as measured by the statewide test. The probability of accurately forecasting mastery of state standards will depend in part on the strength of the relationship between each benchmark assessment and the statewide test (Bergan et al., 2009). Accordingly, it is reasonable to expect significant correlations between benchmark tests in a particular state and the statewide test for that state. A finding revealing such correlations would provide important evidence of the validity of the benchmark assessments (Bergan et al., 2009).

When benchmark goals are aligned with state standards, the ability to accurately forecast the number of students likely to meet state standards is often markedly enhanced. However, alignment is only one issue that needs to be addressed in a forecasting initiative. The accuracy of

benchmark forecasts regarding which students are likely to meet state standards based on benchmark test performance is affected by several factors in addition to benchmark alignment. These factors include the reliability of the benchmark instruments, benchmark validity assessed by the magnitude of the relationship between benchmark tests and the statewide test, changes in standards mastery cut points instituted at the state level, changes in the statewide test, and changes in the benchmark tests. Given the number of variables that may affect forecasting precision, variations in forecasting accuracy are to be expected (Bergan et al., 2009). Additionally, a benchmark reading assessment may be valid for identifying students likely to fall short of proficiency on a state test but may have little validity for diagnosing the specific causes of students' reading difficulties (Herman et al., 2010).

Using student-level data rather than school-level data might help researchers examine the impact of benchmark assessments on important No Child Left Behind subgroups. Another useful follow-up would be disaggregating the school achievement data by subject area and content strand to see if there are any effects in particular standards (Henderson et al., 2007). Higher mathematics scores will come not because benchmarks exist but because of how a school's teachers and leaders use the assessment data. This kind of follow-up research, though difficult, is imperative to better understand the impact of benchmark assessments. A possible approach is to examine initial district progress reports for insight into school buy-in to the initiative, quality of leadership, challenges to implementation, particular standards that participating districts focus on, and how the schools use benchmark data (Henderson et al., 2007).

However, previous research has found that students who do well on one set of standardized tests do not perform as well on other measures of the same content, suggesting that

they have not acquired a deep understanding (Olson, 2005). The core problem lies in the false, but nevertheless widespread, assumption that a benchmark assessment is a particular kind of measurement instrument rather than a process that is fundamental and indigenous to the practice of teaching and learning (Heritage, 2010). Evidently, benchmark assessments are gaining ground in the terms of their perceived, though empirically undocumented, significance for increasing achievement (Heritage, 2010). Much of the rhetoric around benchmark assessments paints a rosy picture. Supporters argue that these tests provide data on student understanding and that teachers' analysis of this data will in turn lead to greater differentiation of instruction and better teaching of content. As a result, student learning will improve. Very little research exists, however, on how benchmark assessments are actually used by individual teachers in classrooms, principals, and districts (Goertz et al., 2009).

Educators used to be convinced that enhanced formative and summative assessments such as benchmark exams would produce gains in student achievement when measured in such narrow terms as scores on state-mandated tests; however, educators are now not convinced (Black et al., 2004). Limiting the ability to definitively make this assertion is the fact that the majority of current research tends to focus on individual assessments and not on the relationship among assessments. Assessment research should examine benchmark assessment use in the context of the broader system of assessment programs. In other words, there remains a need to examine the degree to which assessments of different types inform each other (Goertz et al., 2009). This study aims to fill that gap in the literature by determining whether or not there is a positive correlation between benchmark exam use in California public middle schools and student performance on the end-of-year standardized test, such as the CST.

Summary

The literature discussed several aspects of the ongoing debate on the use of benchmark exams, including their genesis, rationale, perceived value, and connection to standardized tests. Proponents argue that the high stakes placed upon schools by the No Child Left Behind Act of 2002 justifies the use of benchmark assessments in order to better prepare students for the end-of-year state exams. Schools and districts make the claim that student performance on benchmark exams can serve as an accurate predictor of student success on end-of-year state exams. It is viewed as an accountability tool to ensure districts, schools, and classroom teachers are adhering to the curricular mandates set by the federal and state governments. Much is at stake for all educational stakeholders as the financial fates of schools and, in some cases, the compensation of teachers are directly tied to test results. From an instructional perspective, advocates of benchmark exams tout their ability to generate cyclical data that informs and improves instructional practices, allowing for timely revisions and interventions. Through differentiated instruction that is better customized to individual areas of strength and weakness, student achievement gains should occur as a result. This information can also be used in districts, schools, and individual classrooms to dictate and justify resource allocation, such as support staff, professional development, and materials and supplies. Benchmark exams can also be used as pacing guides to facilitate teacher collaboration and planning, ideally resulting in all students having been taught what they need to know by the appropriate deadlines.

Opponents of benchmark exams, and oftentimes of standardized assessment in general, believe there is no appropriate justification or utilization and find them to be another blow to students' fragile states as a result of testing fatigue. They argue that this type exam only assesses the lowest level of student learning, a point contextualized in the theoretical framework of

Bloom's cognitive taxonomy. Due to the inability of a multiple choice or true/false "bubble in" exam to facilitate the calling upon of higher-order thinking skills, preparation for these assessments is generally relegated to rote memorization and regurgitation of basic facts, which is what disdainful educational stakeholders would refer to as teaching to the test. Additionally, it has been alleged that standardized exams are culturally biased and unable to take into account specific needs of various demographics. This argument is strengthened by the disproportionately higher utilization of benchmark exams in urban school districts. In looking at standardized and benchmark assessment through the lens of the theoretical framework of Vygotsky, the lack of opportunity for meaningful feedback and social interaction is troubling and lends to the argument that these are tools to assess low-level skills, not the skills of the 21st century. Whereas advocates praise the concept of benchmark exams and standardized test scores being used as the basis for merit pay, opponents assail this as punitive and unfair, continuing the vicious cycle of the low performing students and schools receiving the fewest financial resources. There is very little arguing the expense of these assessment programs, both financially and with regards to instructional time being supplanted with testing time. Billions of dollars are spent each year on inconsistently and optionally implemented assessment programs that leave educational stakeholders feeling the emphasis on learning has shifted from breadth to depth, particularly in California's case. Teaching has become a race against the testing clock where quantity has far exceeded the importance of quality.

As evidenced by a review of the literature, there is no shortage of opinions, many of them quite passionate, when it comes to benchmark assessments. This is an ongoing topic of heated debate stoked by government mandates, a high stakes educational landscape, and countless publishers and consulting firms trying to capitalize on both of these variables. Each of these

corporations stands to make millions of dollars on what they market to desperate schools districts as a silver bullet. And, with the consequences of failing to meet the government mandates so severe, many of these school districts are buying assessment programs that have not yet proven to improve student achievement on end-of-year state standardized tests. Although there is research available to indicate whether or not benchmark exams or other forms of common assessment can serve as accurate predictors of student performance on end-of-year standardized tests, there has been little studied on the direct impact benchmark exams have on standardized test scores. What little research found on this topic was conducted in other states, so the current study provides a unique opportunity to determine whether public middle schools in California who give benchmark exams perform better on the CSTs than those who do not utilize such common assessments. And, given the fact that the state of California does not yet mandate benchmark exam usage, implementation of a benchmark assessment program looks different from school district to school district. Therefore, the impact of utilization variables such as for how long a school has used the exams, how frequently they are administered, how they are created, whether or not they are mandatory at the school, and teacher satisfaction with the exams will also be analyzed to determine whether there is a correlation between the utilization of benchmark exams and state standardized test performance.

CHAPTER THREE: METHODS

Design

This quantitative, non-experimental, correlational research study determined whether there is a correlation between how California public middle schools use benchmark exams and their students' performance on end-of-year standardized tests. As Gall et al., (2007) states, "Correlational research refers to studies in which the purpose is to discover relationships between variables through the use of correlational statistics" (p. 332). "The basic design in correlational research is very simple, involving nothing more than collecting data on two or more variables for each individual in a sample and computing a correlation coefficient" (Gall, Gall & Borg, 2007, p. 335). Correlational statistics are frequently used in test construction and test analysis; therefore, this was the appropriate selection for this study as it identified the effect the predictor variables of benchmark exam utilization, how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams, had on standardized test scores. Specifically looking at results from the California Standards Test and the state accountability measurement tool, the Academic Performance Index (API), the design identified if there was a direct correlation between how benchmark exams were utilized and the criterion variable of the 2012 API scores.

Research Question

This research question for this study is:

RQ1: Can changes in the API scores of California public middle schools that utilize

benchmark exams be predicted from a linear combination of how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams?

Null Hypothesis

The null hypothesis for this study is:

$H_0$1: Changes in the API scores of California public middle schools that utilize benchmark exams cannot be predicted from a linear combination of how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams.

Participants and Setting

The population for this study was California public middle schools that utilized benchmark exams in some capacity. These schools, by definition, are free of charge, state-funded, co-educational, secular, and offer grades six through eight (Ed-Data, 2014). They were a balanced representation of high-performing (800+), mid-performing (700-799), and low-performing (below 700) schools as determined by their 2011 API scores. For this study, the sample size was 93 schools. According to Gall et al. (2007), in correlational research, a minimum of 30 participants is desirable. A sample size of 66 is the required minimum for a medium effect size with a statistical power of .7 at the .05 alpha level (Gall et al., 2007, p. 145). Schools were chosen using convenience sampling among all California public middle schools. The researcher sent an email to every California public middle school in the state. Included in the email were the informed consent form and survey. Completion of the survey was entirely voluntary. Schools that agreed to be part of the research study completed the survey and returned it to the researcher via the provided link. The researcher stopped collecting survey data

after having received responses from 100 schools. Noticing that 93 of the 100 responding schools utilized benchmark exams in some capacity, the researcher chose to use those 93 schools as the sample.

As the research study was conducted, the setting did not need to be the campuses of participating schools. Instead, research and analysis took place at home in an online environment, using email, SPSS, and the California Department of Education website.

Instrumentation

The instrument used in this research study was an original online survey created by the researcher and had never been used before. The survey was given the same title as the research study itself and was made available using the Survey Monkey website (see Appendix A). The purpose of this instrument was to first ascertain if participating California public middle schools utilized benchmark exams in some capacity. If the school answered affirmatively to using benchmark exams, the survey then specifically asked for information regarding how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and how satisfied the respondents are with the exams. This instrument was developed to obtain the necessary information to analyze if and to what degree a positive correlation existed between how California public middle schools were utilizing benchmark exams and standardized test scores, specifically their performance on the California Standards Tests.

The survey instrument consisted of eight questions, seven of which were multiple choice. The first question asked the name of the participating school in order to obtain current and subsequent year API scores. Included in the first question was a disclaimer stating that school identities would be removed upon aggregation of the collected data. Question two asked if the

school utilizes benchmark exams. If the respondent indicated a negative answer to that question, they responded to the next six questions by selecting "NA." If the respondent indicated the school does utilize benchmark exams in some capacity, they responded to the next six questions with information for the researcher in regards to the predictor variables of the study, such as how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and how satisfied they are with the exams.

The scales of measurement varied from question to question but remained consistent in asking the respondent to select the answer(s) that best describe how the school was currently utilizing benchmark exams. The answer choices for questions two through eight ranged from two possible answers to seven. Scoring was performed by the researcher using the Survey Monkey website and, as there was no specific point value associated with each answer choice, the scoring procedure was an aggregation of data to be entered into the SPSS data sheet. The approximate time to complete the survey was only one minute, as the information required to answer the questions would be common knowledge for the respondent.

Procedures

An email was sent to every public middle school in California asking permission for their participation as outlined in the approved IRB application. IRB approval was granted with an exemption falling under category 46.101 (b)(2,4), which identifies specific situations in which human participants research is exempt from the policy set forth in 45 CFR 46 (see Appendix B). The informed consent form, including a brief explanation of the research study and a link to an electronic survey were included in the email (see Appendix C). Schools willing to participate were asked whether or not they administer benchmark exams, how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the

exams are mandatory, and how satisfied they are with the exams. Responses of participating schools were automatically returned to the researcher using the Survey Monkey program.

An SPPS data spreadsheet was created. The names of all participating schools that utilized benchmark exams in some capacity were listed. The 2011 base API, as obtained from the California Department of Education website, was entered for all schools. Survey response information on predictor variables – how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams – was entered into the SPSS data sheet using numerical codes representing each response. All participating schools' 2012 growth API scores were then entered into the table for criterion variable analysis, i.e. score change. At this time, the names of the participating schools were deleted.

Data Analysis

Data was analyzed through the use of a linear multiple regression analysis. This research problem, like many in education, involved interrelationships between three or more variables. Multivariate statistics, such as multiple regression, allow researchers to measure and study the degree of relationship among various combination of these variables (Gall et al., 2007, p. 352-353). Analysis of the data determined whether or not, and to what degree there was a correlation between the predictor variables of how benchmark exams are utilized – specifically how long the benchmark exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams – and the criterion variable of API score change between the 2011 base and 2012 growth API scores of the schools as calculated by the correlation coefficient.

Data screening was performed on each of the variables (API score change, how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams) in regard to data inconsistencies. Box and whisker plots were utilized to detect outliers on each of the predictor and criterion variables. Normality was analyzed utilizing the Kolmogorov-Smirnov test. The linear multiple regression analysis was performed to assess the null hypothesis utilizing a .05 alpha level. Correlation coefficients showed a low effect size across all variables.

CHAPTER FOUR: FINDINGS

Research Question

The research question for this study was:

RQ1: Can changes in the API scores of California public middle schools that utilize benchmark exams be predicted from a linear combination of how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams?

Null Hypothesis

The null hypothesis for this study is:

$H_0$1: Changes in the API scores of California public middle schools that utilize benchmark exams cannot be predicted from a linear combination of how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams.

Descriptive Statistics

Data collected for the variables of API score change, for how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams can be found in Table 1.

Table 1

Descriptive Statistics

	M	SD	N
Score Change	10.849	21.083	93
How Long	2.462	.6178	93
How Often	3.462	.6685	93
How Created	2.139	1.138	93
Mandatory	1.978	.1458	93
Satisfaction	3.204	.759	93

Results

Data screening

Data screening was performed on each of the variables (API score change, for how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams) in regard to data inconsistencies. No data inconsistencies were found. Box and whisker plots were utilized to detect outliers on each of the predictor and criterion variables (see Figures 1-6). Outliers were discovered in the mandatory and satisfaction data sets. However, the researcher chose to maintain the outliers to obtain a larger sample size. Normality was analyzed utilizing the Kolmogorov-Smirnov test, and violations of normality were found on all of the variables. Knowing this, the researcher understood that the multiple linear regression statistical power was greatly reduced; however, the analysis was continued.

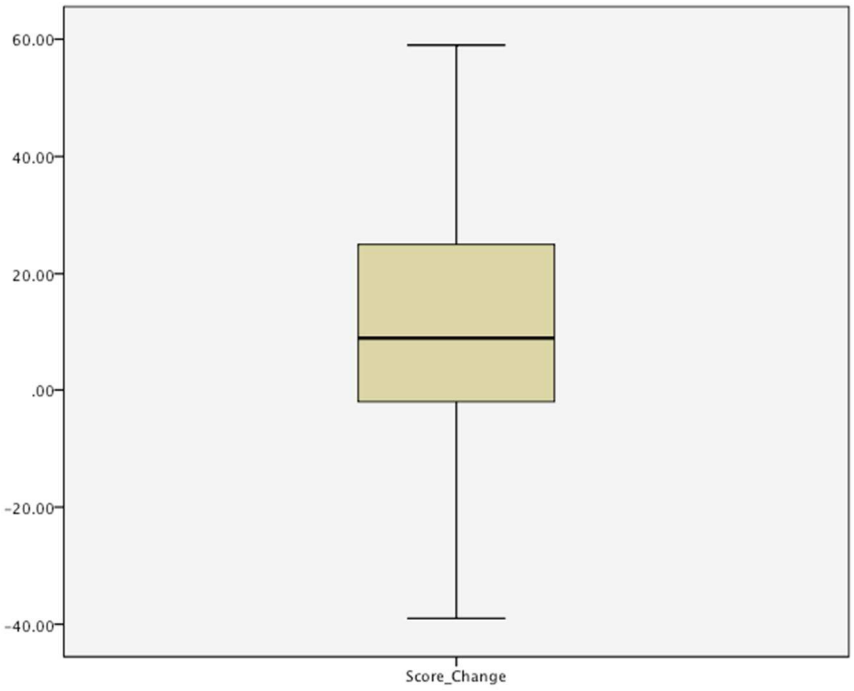

Figure 1.

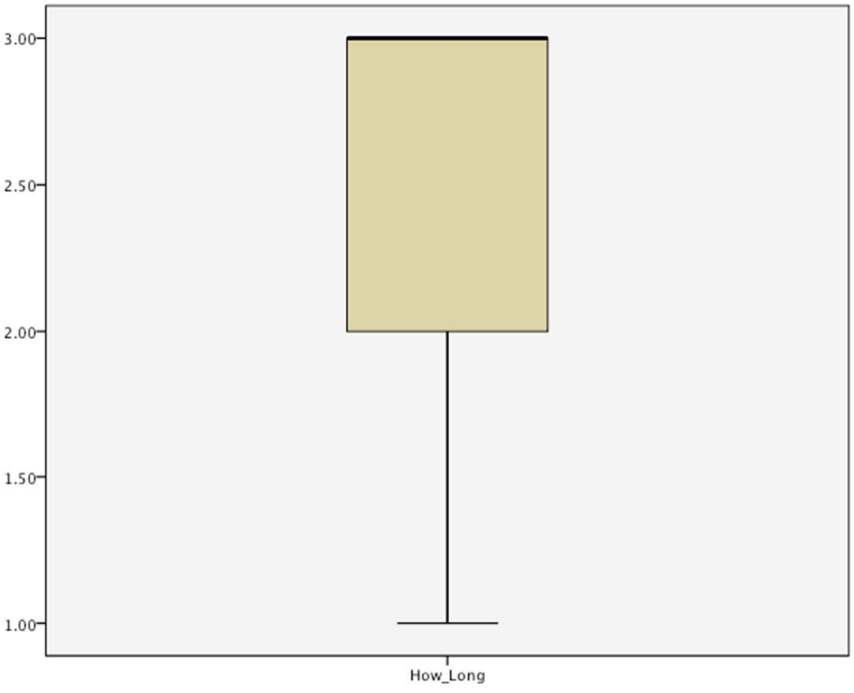

Figure 2.

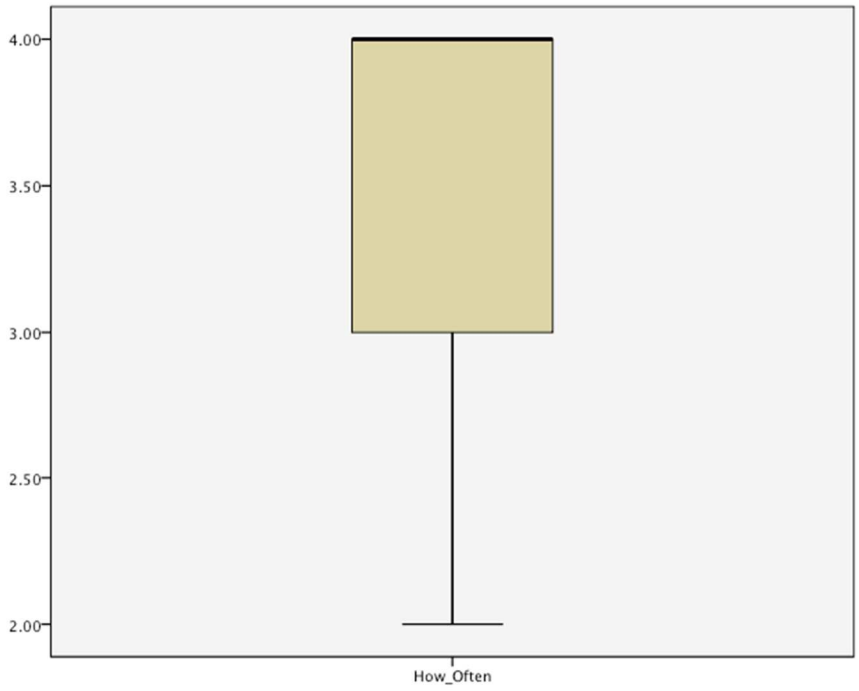

Figure 3.

Figure 4.

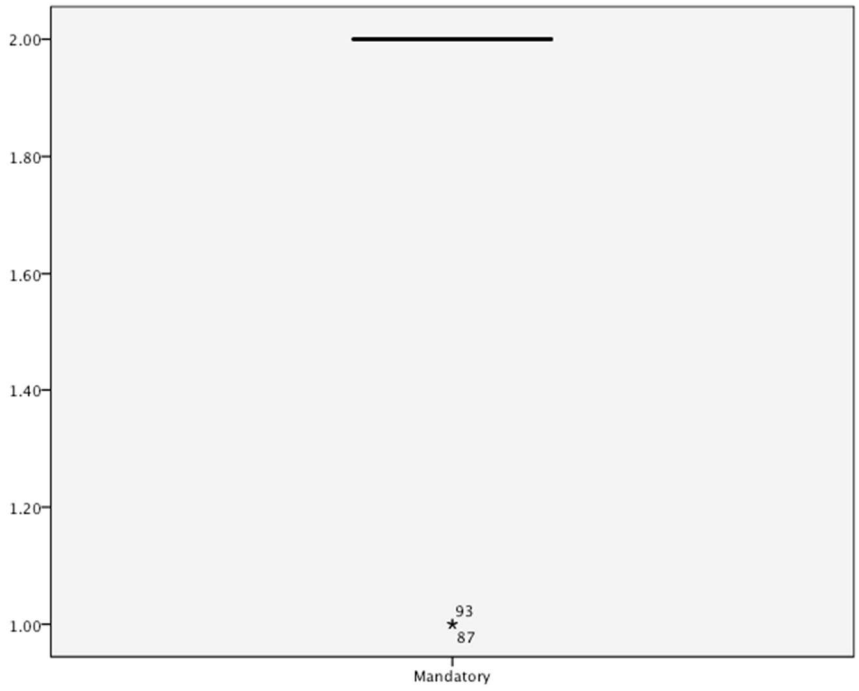

Figure 5.

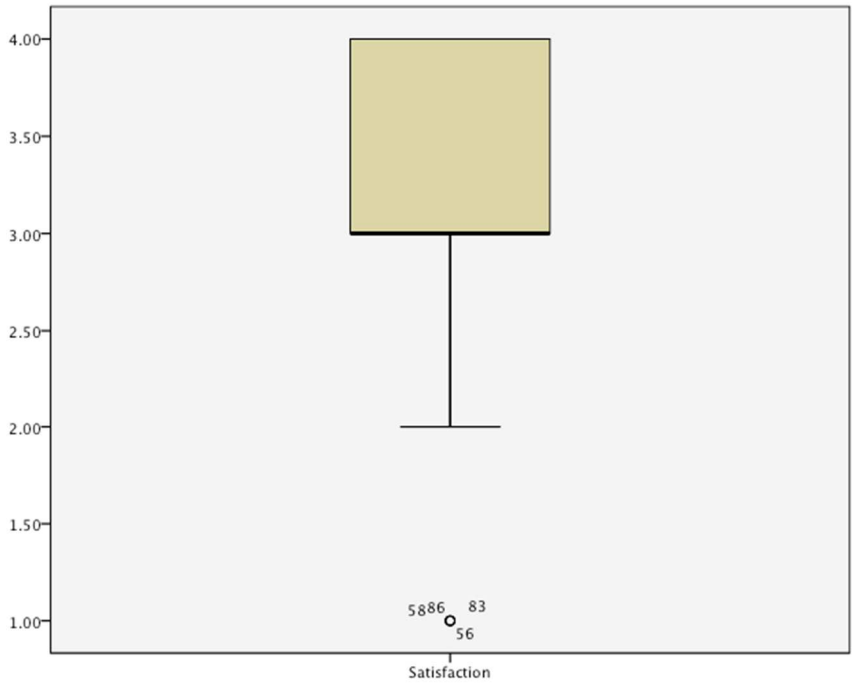

Figure 6.

Null Hypothesis

Results of the statistical analysis. A linear multiple regression analysis was performed to assess the null hypothesis utilizing a .05 alpha level. The researcher failed to reject the null hypothesis $F(5, 87) = 1.35$, $p = .25$. The multiple correlation coefficient for the prediction model was $R = .27$, adj $R^2 = .02$, $R^2 = .07$, meaning that only 7% of the variance can be accounted for by the linear combination of measures. There were no significant contributions among the predictor for how long the exams have been utilized ($p = .88$), how often the exams are administered ($p = .22$), how the exams were created ($p = .10$), whether or not the exams are mandatory ($p = .08$), and teacher satisfaction with the exams ($p = .66$). Tables 2-4 include information about the model summary, ANOVA, and coefficients.

Table 2

Model Summary

Model	*R*	R^2	Adjusted R^2	Std. Error of the Estimate
1	.269	.072	.019	20.884

Note. Predictors: (Constant), Satisfaction, How Created, How Long, Mandatory, How Often

Table 3

ANOVA Test Results

Model	Sum of Squares	df	Mean Square	F	Sig
Regression	2949.853	5	589.971	1.353	.250a
Residual	37946.040	87	436.161		
Total	40895.892	92			

Notes. Criterion variable: score change.

[a] = Predictors: (constant), satisfaction

Table 4

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Significance
	B	Std Error	Beta		
1 (Constant)	37.799	31.658		1.194	.236
How long	.563	3.363	.016	.155	.877
How often	4.256	3.415	.135	1.246	.216
How created	3.221	1.943	.174	1.658	.101
Mandatory	-27.311	15.213	-.189	-1.795	.076
Satisfaction	1.270	2.913	.046	.436	.664

Note. Criterion variable = score change.

CHAPTER FIVE: DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

Discussion

The purpose of this study was to determine whether or not there is a direct relationship between how benchmark exams are utilized and end-of-year standardized test scores, particularly in public middle school campuses in California. The null hypothesis stated that changes in the API scores of California public middle schools that utilize benchmark exams cannot be predicted from a linear combination of how long the exams have been utilized, how often the exams are administered, how the exams were created, whether or not the exams are mandatory, and teacher satisfaction with the exams. The researcher failed to reject the null hypothesis. In other words, there was no significant linear relationship between any of the administrative predictor variables and student performance on the end of the year standardized tests. Therefore, the API score changes between the 2011 base scores and the 2012 growth scores were unrelated to how the benchmark exams were administered.

The hypothesis was restated and discussed in light of the results, literature, other studies, and theory. The discussion looked at whether the results support or contradict other studies and theory. Although neither based in California nor specific to public middle schools, previous studies such as Sherman's (2008) dissertation on the Texas state-mandated benchmark exams and Hefflin's (2009) dissertation on the Pennsylvania state-mandated 4Sight benchmark assessments attempted to determine the extent to which benchmark exams serve as predictors on standardized tests. The results of this study neither support nor contradict their findings, as the impact benchmark exam utilization had on student standardized test scores was not determined. The most comparable study took place in Massachusetts when a research team used a quasi-experimental design to examine the effectiveness of a benchmark exam pilot program. The

study intended to identify whether or not schools using the benchmark exams showed greater gains in student achievement than schools not in the program. After two years of program implementation, no statistically significant difference in test scores could be found between schools participating in the benchmark assessment pilot program and the comparison schools. The results of this study support those findings, although both studies reflected similar limitations in the data.

The theoretical frameworks of Vygotsky (1978) and Bloom et al. (1956) are also supported by the results of this study. The absence of a significant positive linear relationship between how benchmark exams were utilized and student performance on end-of-year standardized tests supports the assertion that this type of exam can only assess low-level student learning and deprives students the opportunity for the meaningful feedback and interactions necessary to develop higher order thinking skills.

Conclusion

Since the inception of the State Testing and Reporting program in 1998, educators in California public middle schools and beyond have tirelessly searched for a strategy that statistically improves student performance on end-of-year standardized tests (Ed-Data, 2014). In California, the California Standards Test (CST) administered each May, is used to evaluate schools and districts by assigning them a score based on student results in four subject areas: English, math, science, and history. This Academic Performance Index (API) was not only perceived by stakeholders as an evaluative mark, but also dictates whether or not the school and/or district is subjected to corrective actions taken by the state.

The utilization of benchmark exams, voluntarily implemented by a large number of schools and districts, was thought to be such a strategy. As not yet mandated by the California

Department of Education, whether or not benchmark exams are utilized at all, along with every aspect of their logistical implementation has been left to the individual schools and districts. In surveying 100 California public middle schools, the data shows that the overwhelming majority of schools are, in fact, utilizing benchmark exams in some capacity. What remained unclear was whether or not the administration of the benchmark exams could be positively correlated with increased student performance on the end-of-year standardized tests.

As the literature confirms, there is an abundance of fierce and deeply rooted beliefs about benchmark assessments. Federal and state government involvement adds fuel to the fire of this continuing discussion as does the high stakes educational landscape we live in. Meanwhile, an abundance of publishers and consulting firms are eagerly trying to take advantage of both of these conditions. Each of these businesses stands to make huge sums of money on what they market to desperate schools districts as a quick fix. Desperate to avoid the penalties levied upon schools and districts who fail to meet these government mandates, unproven assessment programs are being purchased at an alarming rate with little to no statistical evidence supporting their claims to improve student achievement on end-of-year state standardized tests. Whereas there is plenty of research available indicating whether or not benchmark exams or other forms of common assessment can accurately predict student performance on end-of-year standardized tests, little has been studied on the direct impact benchmark exam utilization has on standardized test scores. The limited research on this topic was conducted outside of California; therefore, the study at hand creates a unique opportunity to identify whether public middle schools in California who give benchmark exams perform better on the CSTs than those who do not utilize such common assessments. Another way the existing body of literature falls short is in clearly identifying whether or not there is a specific prescription of administration variables that can

make benchmark exams effective enough to directly and positively impact student performance on the end-of-year standardized tests. As the state of California has not yet mandated benchmark exam usage, implementation of benchmark assessment programs look different from school district to school district. As a result, the significance of administration variables – for how long a school has used the exams, how frequently they are administered, how they are created, whether or not they are mandatory at the school, and teacher satisfaction with the exams – will be analyzed to assess the degree of correlation between the utilization of benchmark exams and state standardized test performance.

The study highlights the fact that there remains yet to be a silver bullet proven to positively impact student performance on standardized tests. In fact, given the absence of a positive correlation, one might wonder and perhaps research whether or not the presence of an additional battery of testing throughout the year negatively impacts student performance on standardized test scores. There are several detractors of benchmark exams who cite test fatigue, teachers teaching to the tests, a curricular favoring of breadth versus depth, and inauthentic assessment as their bases of animosity. When considering the opposing argument rationalized such exams by claiming it would all be worth it in the end once the end-of-year exam scores arrived, it could be concluded that the results of this research study supports the detractors.

As the momentum against end-of-year high-stakes exams increases, some districts and entire states are taking drastic measures by legislating the right of families to opt-out of the tests. Not surprisingly and, in a decision somewhat validated by the results of this study, California was the first state to implement opt-out procedures. California Education Code Section 60615 states, "Notwithstanding any provision of law, a parent's or guardian's written request to school officials to excuse his or her child from any or all parts of the assessments administered pursuant

to this chapter shall be granted" (Opt Out of Standardized Tests, 2015). Although once considered a rarity, the opt-out push has prompted high-profile boycott efforts and meetings in large districts such as Chicago and led more parents nationwide to join forces with anti-testing advocates in arguing that the assessments are unnecessary, excessive, and, in some cases, even harmful to students (Education Week, 2014). This is a legal battle that will continue to rage, not only in California but nationwide. And, for every research study, such as this one, that fails to support the benefit of standardized testing and the utilization of accompanying batteries of benchmark exams, the tide will continue to turn against these practices.

Implications

As the study failed to indicate a significant, positive, linear relationship between how benchmark exams are utilized and student performance on end-of-year standardized tests, schools and school districts continue to be without significant verification that there is a particular prescription of administrative variables shown to yield positive results. Stakes continue to rise and school districts, schools, administration, and teachers all find themselves under increasingly immense pressure. Additionally, in an era of budget cuts, increasing class sizes, and decreasing resources, finances must be protected and well utilized. Therefore, studies such as this could potentially influence the educational landscape of California by validating its refusal to mandate benchmark exams. Additionally, as less value is placed upon yearlong batteries of assessments in the name of standardized test scores, it is the converse, negative impact of these assessments that has led California and other states and districts to defy federal mandates by allowing families to opt out.

When considering the millions of dollars spent by tax-funded California public school districts on the creation and administration of benchmark exams in an attempt at best preparing

students for end-of-year standardized tests, the implications of this study and similar studies are not simply pedagogical but also financial. Because the results of this study indicated the administrative predictor variables had no significant impact on the criterion variable of student performance on end-of-year standardized tests, the evidence supporting the gross financial expenditures on benchmark exams continues to be elusive. This study highlighted the fact that there is no direct correlation between how benchmark exams are utilized and standardized test scores, which should considered by authorities controlling the educational purse strings of schools, districts, and states. In addition to the assertion that benchmark testing may very well be a waste of money, one must also consider the amount of hours of sacrificed instructional time in exchange for a perceived more comprehensive testing program. Opponents have argued for years that classroom time would be far better spent authentically teaching and learning, rather than training for and administering additional standardized tests. The results of this study support that sentiment and call into question the unjustified expenditure of these ongoing, yearlong assessment batteries that have yet to be proven effective in the single task they are designed to accomplish.

Limitations

Limitations of the study include its narrow scope of only including California public middle schools. Student attendance, motivation, health, discipline, teacher quality, administrative leadership, and countless other factors influence standardized test results. Although the researcher will attempt to counter for these limitations through a balanced representation of schools, those with high percentages of second language learners, socioeconomically disadvantaged families, and students designated as having special education needs may perform lower than schools with large percentages of historically high achieving

demographics such as Caucasian and Asian students (California Department of Education, 2015).

Outliers discovered in the mandatory and satisfaction data sets threaten validity. However, the researcher chose to maintain the outliers to obtain a larger sample size. Normality was analyzed utilizing the Kolmogorov-Smirnov test, and violations of normality were found on all of the variables. Knowing this, the researcher understood that the multiple linear regression statistical power was greatly reduced; however, the analysis was continued.

Recommendations for Future Research

Given the limited scope of the study and failure to reject the null hypothesis, the researcher recommends greatly expanding the sample size by including California public elementary and high schools. The goal would be to include enough participating schools that the percentage of schools not utilizing benchmark exams at each grade level would be comparable to those that do. As this study failed to show a positive correlation between how benchmark exams are utilized and student performance on standardized tests, the researcher recommends conducting a study to identify whether or not there is a positive correlation between simply utilizing benchmark exams and student performance on standardized tests, regardless of how they are utilized. Another approach would be to disaggregate data by CST subject area or content strand in order to see if benchmark exams are more effective at preparing students for certain subject area portions of the CST. For example, educational stakeholders may find that it is easier to create a benchmark exam that closely mimics the mathematics portion of the CST than it is to do so with the English portion of the exams. This additional layer of analysis could identify whether or not additional testing batteries, such as benchmark exams, increase student achievement on certain subject areas of the end-of-year exams, but not others.

The researcher would also recommend more closely examining the effect of mandatory predictor variable. In utilizing the .05 alpha level, this variable's *p*-value of .08 indicates a lack of significance although, when compared to the *p*-values of the other four predictor variables, it shows a considerably higher level of significance. Adjusting the level of confidence from 95% to 92% would indicate that, whether or not the benchmark exams were mandatory was, in fact, a significant contributor. Further study and data analysis could possibly assert that, if the use of benchmark exams is mandated by site, district, or state administration, their direct impact on end-of-year standardized test score is significant.

In reference to the argument that benchmark exams are a waste of time and money, an argument supported by the findings of this study, the researcher recommends further study of the perceptions of teachers regarding benchmark exam utilization. Whereas the satisfaction predictor variable was based on a four-point scale, it did not identify why, specifically, some teachers may be more satisfied with their benchmark program than others. Perhaps the amount of lost instructional time significantly affects teachers' attitudes towards the value of these exams. Also likely, the amount of lost instructional time is directed correlated to student performance on these and other exams.

In conclusion, the researcher's failure to reject the null hypothesis of this study lends itself to further exploration of the correlation between benchmark exams and student performance on end-of-year standardized tests. Results of future research, particularly on the above-mentioned areas, may support the findings of this study and provide further justification for educational agencies to abandon benchmark exam programs. There would certainly be a large contingent of educators claiming they knew this all along. Or, perhaps, there are additional

variables, or combinations thereof, regarding benchmark exam utilization that significantly and directly impact student performance on end-of-year standardized tests.

REFERENCES

Abrams, L., McMillan, J. & Wetzel, A. (2010). *Teachers' voices about the effectiveness of Benchmark testing.* Virginia Commonwealth University. Williamsburg, VA.

Action Learning Systems, Inc. (2004). English-language arts benchmark tests. Technical Report.

Akdemir, O. & Colakoglu, O. (2010). Motivational measure of the instruction compared: Instruction based on the ARCS motivation theory vs. traditional instruction in blended courses. *Turkish Online Journal of Distance Education, 11*(2), 73-89.

ASCD. (2005). Retrieved from www.ascd.org/publications/educationalleadership/nov05/vol63/num03/Making-Benchmark-Testing-Work.aspx.

Bergan, J.R., Bergan J. R. & Burnham, C.G. (2009). Benchmark assessments in standards-based education. The Galileo K-12 Online Educational Management System. Assessment Technology, Incorporated. Tucson, AZ.

Bigelow, B. (1999). Why standardized tests threaten multiculturalism. *Educational Leadership, 56(*7).

Black, P., Harrison, C., Marshall, B. & William, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 80*(2), 9-21.

Bloom, B., Englehart, M. Furst, E., Hill, W. & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals.* Handbook I: Cognitive domain. New York, Toronto: Longmans, Green.

Butler, F. & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: Current trends and old dilemmas. *Language Testing, 18*(4), 409-427.

California Department of Education. (2011). Retrieved from www.cde.ca.gov

Coffey, H. (2009). Benchmark assessments. Retrieved from www.learnnc.org

Darling-Hammond, L. & Adamson, F. (2013). *Developing assessments of deeper learning: The costs and benefits of using tests that help students learn.* Stanford, CA: Stanford University Center for Opportunity Policy in Education.

Ed-Data. (2014). Retrieved from www.ed-data.k12.ca.us

Ed-Evidence. (2009). What are benchmark assessments and how to they work? Retrieved from www.archive.relnei.org/newsletters

Education Week. (2014). Retrieved from www.edweek.org

Evans, J. (2013). Problems with standardized testing. *Education.com*. Retrieved from www.education.com/reference/article/Ref_Test_Problems_Seven/

FairTest: The National Center for Fair and Open Testing. (2009). *The value of formative assessment.* Retrieved from www.fairtest.org

Furgeson, J., Gill, B., Haimson, J., Killewald, A., McCullough, M., Nichols-Barrer, I., The, B. & Verbitsky-Savitz., N. (2012). *Charter school management organizations: Diverse strategies and diverse student impacts.* Cambridge, MA: Mathematica Policy Research

Gall, M., Gall, J. & Borg, W. (2007). *Educational research: An introduction*. Boston: Pearson Education.

Goertz, M.E., Olah, L.N. & Riggan, M. (2009). *Can interim assessments be used for instruction change?* (CPRE Policy Brief RB-51). Philadelphia, PA: Consortium for Policy Research in Education.

Huag, J. (2010). Teachers won't be evaluated based on test scores. *Las Vegas Review Journal*, B5.

Heffli, P. (2009). *Do benchmark assessments increase student achievement on state standardized tests?* (Unpublished doctoral dissertation). Duquesne University, Pittsburg,Pennsylvania).

Henderson, S., Petrosino, A., Guckenburg, S. & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievemen*t (Issues & Answers Report, REL 2007 – No. 039). Washington, DC: US Department of Education, Institute of Educational Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs.

Heppen, J., Jones, W., Faria, A., Sawyer, K., Lewis, S., Horwitz, A., Simon, C., Uzzell, R. & Casserly, M. (2011). *Using data to improve instruction in the Great City Schools:Documenting current practice*. American Institutes for Research and Council of the Great City Schools.

Heritage, M. (2010). Formative assessment and next-generation assessment systems: Are we losing an opportunity? Washington, D.C.: Council of Chief State School Officers.

Herman, J.L. (2009). *Moving to the next generation of standards for science: Building on recent practices* (CRESST Report 762). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J.L. & Baker, E.L. (2005). Making benchmark testing work. *Educational Leadership*,*63*(3).

Herman, J.L., Osmundson, E. & Dietel, R. (2010). *Benchmark assessment for improved learning* (AACC Report). Los Angeles, CA: University of California.

Kastenbaum, S. (2012). The high stakes of standardized tests. *CNN Radio*. Retrieved from www.schoolsofthought.blogs.cnn.com

Keller, J. (1983). Motivational design of instruction. In Reigeluth, C. M. (Ed.), *Instructional-design theories and models: An overview of their current status*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McNeil, M. (2008). Benchmarks momentum on increase: Governors' group, state chiefs eyeing international yardsticks. *Education Week, 27*(27), 1.

Nelson, H. (2013). Testing more, teaching less: What America's obsession with student testing costs in money and lost instructional time. American Federation of Teachers.

North Dakota Department of Public Instruction. (2015). Retrieved from: www.dpi.state.nd.us/title1/progress/terms.shtm.

Norwalk-La Mirada Unified School District. (2015). Retrieved from www.nlmusd.k12.ca.us.

Oceanside Unified School District. (2015). Program improvement: No Child Left Behind (NCLB) Act of 2001. Retrieved from www.oside.k12.ca.us/nclbpi.

Olson, L. (2005a). Benchmark assessments offer regular achievement. *Education Week, 25*(13).

Olson, L. (2005b). Not all teachers keen on periodic tests. *Education Week, 25*(13), 13.

Opt Out of Standardized Tests. (2015). Retrieved from:www.optoutofstandardizedtests.wikispaces.com.

Salpeter, J. (2004). Data: Mining with a mission. *Technology and Learning, 24*(8), 30-37.

Sawchuk, S. (2009). In funding common assessments, tough challenges: Officials gathering input on best way to leverage $350 million for initiative. *Education Week, 29*(12), 16.

Standardized Testing and Reporting Program. (2014). Retrieved from www.startest.org.

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta* Kappan International, 83(10), 758-765.

Supovitz, J.A. & Klein, V. (2003). Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement. University of Pennsylvania, Graduate School of Education: Center on Reinventing Public Education.

Taking Center Stage. (2001). Retrieved from www.pubs.cde.ca.gov/tcs.

Taking Center Stage – Act II. (2010). Retrieved from www.pubs.cde.ca.gov/tcsii.

Topol, B., Olson, J., Robert, E. & Hennon, P. (2013). *Getting to higher-quality assessments:* Evaluating costs, benefits, and investment strategies. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

US Department of Education. (2012). Retrieved from www.ed.gov.

US Legal. (2015). Retrieved from www.uslegal.com.

Vygotsky, L.S. (1978). Mind and society: The development of higher mental processes. Cambridge, MA: Harvard University Press.

Zehr, M. (2006). Monthly checkups. *Education Week, 25*(35).

APPENDIX A

Instrument

BENCHMARK
EXAM UTILIZATION IN CALIFORNIA MIDDLE SCHOOLS AND STANDARDIZED TEST SCORES: A NON-EXPERIMENTAL CORRELATIONAL STUDY

*** 1. What is the name of your school and school district? (Note: School identities will be removed upon aggregation of collected data)**

*** 2. Does your school utilize benchmark exams?**

- Yes
- No

*** 3. If yes, for how long has your school been utilizing them?**

- Last year was our first year
- 2-4 years
- 5+ years
- NA - We didn't use benchmarks last year

*** 4. If yes, how often were they administered?**

- 1 time per year
- 2 times per year
- 3 times per year
- 4+ times per year
- NA - We didn't use benchmarks last year

*** 5. If yes, how are the exams created?**

- ○ Teacher created
- ○ Primary source test bank
- ○ Third party consulting firm
- ○ Other
- ○ NA - We didn't use benchmarks last year

*** 6. If yes, how does your school utilize benchmark exam data? (Choose all that apply)**

- ☐ Teachers are provided with the data
- ☐ Data are analyzed in PLCs
- ☐ Data are used to inform/revise instructional practices
- ☐ Data are used to assist in adherence to pacing guides
- ☐ Benchmark data are submitted to the district office
- ☐ Other
- ☐ NA - We didn't use benchmarks last year

*** 7. Is the administration of benchmark exams considered mandatory?**

- ○ Yes, and enforced
- ○ Yes, but not enforced
- ○ No
- ○ NA - We didn't use benchmarks last year

*** 8. How satisfied are you with your benchmark exams and their utilization?**

Very Unsatisfied	Unsatisfied	Somewhat Satisfied	Very Satisfied	N/A - Didn't use benchmarks last year
○	○	○	○	○

APPENDIX B

LIBERTY UNIVERSITY

The Graduate School at Liberty University

October 1, 2012

Michael Marcos
IRB Exemption 1402.100112: Benchmark Exam Utilization in California Middle Schools and Standardized Test Scores: A Non-Experimental Correlational Study

Dear Michael,

The Liberty University Institutional Review Board has reviewed your application in accordance with the Office for Human Research Protections (OHRP) and Food and Drug Administration (FDA) regulations and finds your study to be exempt from further IRB review. This means you may begin your research with the data safeguarding methods mentioned in your approved application, and that no further IRB oversight is required.

Your study falls under exemption category 46.101 (b)(2,4), which identifies specific situations in which human participants research is exempt from the policy set forth in 45 CFR 46:

> (2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), **survey procedures**, interview procedures or observation of public behavior, **unless:**
> **(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects**; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.
>
> (4) Research involving the **collection or study of existing data**, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator **in such a manner that subjects cannot be identified**, directly or through identifiers linked to the subjects.

Please note that this exemption only applies to your current research application, and that any changes to your protocol must be reported to the Liberty IRB for verification of continued exemption status. You may report these changes by submitting a change in protocol form or a new application to the IRB and referencing the above IRB Exemption number.

If you have any questions about this exemption, or need assistance in determining whether possible changes to your protocol would change your exemption status, please email us at irb@liberty.edu.

Sincerely,

Fernando Garzon, Psy.D.
Professor, IRB Chair
Counseling

(434) 592-4054

LIBERTY UNIVERSITY.

Liberty University | Training Champions for Christ since 1971

APPENDIX C

Participant Consent Form

Consent Form

Benchmark Exam Utilization in California Middle Schools and Standardized Test Scores: A Non-Experimental Correlational Study

A research study in partial fulfillment of the requirements for the Doctorate of Education

Michael F. Marcos

Liberty University

School of Education

You are invited to be in a research study of whether or not there is a correlation between the utilization of benchmark exams and Academic Performance Index scores in CA public middle schools. You were selected as a possible participant because you are a public middle school in CA. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by: Michael F. Marcos, Ed.D. candidate, Liberty University – School of Education
Background Information:
The purpose of this study is: The purpose of this quantitative non-experimental correlational research study is to determine whether there is a correlation between the utilization of benchmark exams and standardized test performance. How frequently they are administered will be considered a variable. How the exams are created, i.e. teacher-created, primary source test-bank, or third party consulting firms, will be additional variables allowing for a deeper level of analysis should a positive correlation exist.

Procedures:

If you agree to be in this study, we would ask you to do the following things: Answer the questions regarding your school's utilization of benchmark exams in the brief survey attached in the Survey Monkey link. It should take no more than 5 minutes.
Risks and Benefits of being in the Study:

The study has very minimal risk – no more than the participants would encounter in everyday life. In the highly unlikely event that the names of participating schools were identifiable in the publication of the analysis results, these schools could be placed under scrutiny regarding their utilization of benchmark exams (or lack thereof) and their API scores. It could convince parents of prospective students to select another school and potentially cost the participating school district revenue through loss of average daily attendance (ADA). Therefore, immediately after

receiving the informed consent of the participating schools, their identities will be made completely anonymous as they will be placed in a categorizing group depending on academic performance.

Benefits to participation: Society, particularly California, stands to benefit from this study as it will clearly identify whether or not a tool that some states and schools have made mandatory, truly provides a proven method of increasing student achievement. If such a positive correlation is identified, schools will be more inclined to utilize benchmark exams which could create statistically higher performing schools and ultimately raise neighboring property values. Eventually, this could lead to a CA state mandate or even a national mandate to use benchmark exams in schools. Should this study indicate there is a direct positive correlation between the utilization of benchmark exams and student performance on end-of-year standardized tests, schools and school districts would now have significant verification that there is, in fact, a system proven to yield positive results. If no direct positive correlation between the utilization of benchmark exams and student performance on end-of-year standardized tests is statistically indicated, school districts will have valuable information on which to amend professional development and data collection protocols.

Liberty University will not provide medical treatment or financial compensation if you are injured or become ill as a result of participating in this research project. This does not waive any of your legal rights nor release any claim you might have based on negligence.

Compensation:

You will NOT receive payment for participation in this study.

Confidentiality:

The records of this study will be kept private. In any sort of report I might publish, I will not include any information that will make it possible to identify a subject. Research records will be stored securely and only researchers will have access to the records.

Only the names of the participating middle schools will initially be identifiable to ensure proper grouping. No individual data will be collected whatsoever. Once the data is collected, the analysis and reporting phase will be 100% anonymous.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with Liberty University. If you decide to participate, you are free to not answer any question or withdraw at any time without affecting those relationships.

Contacts and Questions:

The researcher conducting this study is: Michael F. Marcos. You may ask any questions you have now. If you have questions later, **you are encouraged** to contact him at

mfmarcos@liberty.edu . He is conducting this study under the direction of Dr. Jared T. Bigham who may also be contacted with any questions or concerns: jtbigham@liberty.edu .

If you have any questions or concerns regarding this study and would like to talk to someone other than the researcher(s), **you are encouraged** to contact the Institutional Review Board, Dr. Fernando Garzon, Chair, 1971 University Blvd, Suite 1582, Lynchburg, VA 24502 or email at fgarzon@liberty.edu.

You will be given a copy of this information to keep for your records.

Statement of Consent:

I have read and understood the above information. I have asked questions and have received answers. I consent to participate in the study.

Signature: _____ Date: _____

Signature of Investigator:_____ Date: _____

IRB Code Numbers: [ ] (After a study is approved, the IRB code number pertaining to the study should be added here.)

IRB Expiration Date: [ ] (After a study is approved, the expiration date (one year from date of approval) assigned to a study at initial or continuing review should be added. Periodic checks on the current status of consent forms may occur as part of continuing review mandates from the federal regulators.)