UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

# INTERIM ASSESSMENT SYSTEMS AND COGNITIVE MEASURES OF DEEPER LEARNING: AN EVALUATION OF THE USA TESTPREP® PROGRAM

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF EDUCATION

By

LEEDY K. SMITH
Norman, Oklahoma
2018

# INTERIM ASSESSMENT SYSTEMS AND COGNITIVE MEASURES OF DEEPER LEARNING: AN EVALUATION OF THE USA TESTPREP® PROGRAM

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF EDUCATIONAL LEADERSHIP AND POLICY STUDIES

BY

______________________________
Dr. Timothy G. Ford, Chair

______________________________
Dr. Curt Adams

______________________________
Dr. Keith Ballard

______________________________
Dr. Patrick Forsyth

______________________________
Dr. Vickie Lake

*To my dad who taught me who I am, whose I am, and what I am about. You serve continuously, give without measure, and fill our family with boundless joy and laughter. We know our Father because of you.*

*And to my mother, whose beauty and grace is only exceeded by the beauty of her heart. Your kindness and patience bless so many, particularly this strong-willed child.*

*Blessed be the name of the Lord.*

# Acknowledgements

I would like to express my profound gratitude to the members of my dissertation committee, Dr. Timothy G. Ford, Dr. Curt Adams, Dr. Keith Ballard, Dr. Patrick Forsyth and Dr. Vickie Lake, for their wisdom and guidance throughout this journey. Each of you have helped shaped my studies and inspired me along the way. I am grateful. In particular, I would like to mention my committee chair, Dr. Timothy G. Ford. To say that I am appreciative and thankful for your mentorship of me and my work is woefully inadequate. Your feedback and encouragement during this endeavor has made it all possible.

I would be remiss to not acknowledge my family and friends. Thank you for your support and encouragement. You filled in the gap so that I could pursue my dreams and cheered me along the way. I am blessed.

# Table of Contents

## List of Tables

## List of Figures

# Abstract

Recent federal and state policy has placed increasing focus on college and career readiness. As a result, many states and local districts respond by implementing interim assessment systems to help them facilitate instruction and monitor progress towards college and career readiness. Yet, there is little evidence to support the efficacy of these programs in influencing student outcomes (Perie, et al., 2009). Thus, the purpose of this study is to consider to what extent one interim assessment program, USA TestPrep®, as implemented in a suburban Oklahoma middle school met the goals of college and career readiness as examined through a lens of deeper learning. Through a mixed methods approach, the researcher concluded that while the interim assessments, as created by teachers in the district, were predictive of the standardized assessments, they did not align with the measures of deeper learning. As a tool for data driven decision making, teachers' primary perception of the program was one of predictive purpose and to a lesser extent the instructional purposes to inform teaching and learning.

# Chapter 1: Introduction

Over the past decade, the Common Core State Standards (CCSS) movement and the *Race to the Top* initiative have placed considerable priority on college and career readiness (CCR) as a key outcome for students (Conley, Drummond, de Gonzalez, Rooseboom, & Stout, 2011; Lombardi, Conley, Seburn, & Downs, 2012). Even with increased control at the federal and state level in recent years, a substantial proportion of American students remain ill-prepared to meet the demands of college and careers (Conley & Darling-Hammond, 2013; Bitter & Loney, 2015). Student achievement either remains stagnant or in some instances has declined (Amrein & Berliner, 2002; Darling-Hammond, Wilhoit, & Pittenger, 2014) while significant achievement gaps and educational inequalities persist (Darling-Hammond, 2007; Noguera, Darling-Hammond & Friedlaender, 2015). Additionally, the educational system as a whole has not evolved to deliver deeper learning experiences that fit the needs of a modern workforce (National Research Council, 2012).

Some of this stagnation is likely due to the considerable variation across state definitions of college and career readiness, and the standards and assessments they use to measure it (Lee & Reeves, 2012; Linn, Baker & Betebenner, 2002). Kobrin (2007) and Mayurama (2012) note that definitions of college and career readiness at the state and local levels vary substantially, ranging from: a) earning a standard high school diploma and reading at a basic level to b) earning high student grade point average, class rank or college admission test scores to c) success in college level courses or d) a demonstration of deeper learning through measurement of various cognitive and non-cognitive outcomes. In many of these cases, variation in the operationalization of college and career readiness is the result of the absence of a clear definition of the term; it is defined by whatever learning outcomes local and state assessments happen to measure (Adams et al., 2017). Because the definition is a moving target, districts and states are challenged when seeking the alignment of teaching and learning to college and career readiness.

In the recent literature, preparing students to be college and career ready is often associated with a set of competencies defined as deeper learning. Deeper learning is the ability to master content-specific knowledge and then transfer that knowledge to use in new and unique situations (NRC, 2012; Bitter & Loney, 2015; Huberman et al., 2014). The William and Flora Hewlett Foundation (2013) and Bitter and Loney (2015) assert that deeper learning experiences better prepare students for success in college, career, and civic life. Conley and Darling-Hammond (2013) assert that preparing students to be college and career ready necessitates that schools and educational systems move beyond current curriculum, instruction, and assessment practices to instructional systems and structures that are able to support the development of deeper learning.

## Statement of the Problem

Increased attention to improving achievement, deeper learning, and/or college and career readiness has precipitated the development, marketing, and use of interim assessment systems to states and school districts to track progress towards these goals. Many districts and schools, feeling the pressure to improve, adopt these systems in the hopes they will assist teachers in developing students who are college and career ready. These systems contain tools, resources, and information purported to increase student outcomes through increasing instructional capacity and management (Perie, Marion, &

Gong, 2009; Goertz et al., 2009). Yet, Goertz, Oláh, and Riggan (2009) demonstrate that, despite their widespread use, little evidence of the interim of assessment effectiveness exists. Moreover, we know even less about how they are being used to improve instruction and increase learning outcomes including deeper learning and college and career readiness.

## Purpose of the Study

USA TestPrep® is an interim assessment system which is marketed as both an assessment tool and a capacity tool. Its creators claim it can be used for benchmarking achievement and college and career readiness and assisting teachers and other staff in the use of these data for instructional decision making ("Comprehensive Solution," 2017). This system offers a number of training videos and manuals in operating the program and its various features. The program allows for teachers to use program generated benchmark tests, select program-generated questions to create their own benchmark tests, or even input teacher-created questions as part of benchmark testing. Aside from the benchmark tests, activities include video lessons, practice questions, bell work, vocabulary drills, and games. The program content allows for teachers to assign activities as well as for individual students to choose their own assignments and work at their own pace. Each activity and question is identified by standard, objective, and depth of knowledge. Both teachers and students may run progress reports. While the students may run only their own individual progress reports, teachers can consider student mastery at the individual, class, teacher, or school level.

Given these system features, USA TestPrep® seems, at first glance, to be a useful tool for building the capacity of teachers in meeting the learning needs of students. However, prior studies (see, for example, Goertz et al., 2009) of the effectiveness of interim assessments adopted by districts and schools to meet *Race to the Top* goals have revealed little, if any, evidence that they are meeting their stated aims of improving teaching and learning towards college and career readiness. The purpose of this study is to examine the effectiveness of the USA TestPrep® system, an exemplar of an interim assessment program targeting deeper learning and college and career outcomes, by gathering empirical evidence of its utility in helping districts and teachers facilitate these experiences and outcomes on the part of students. To this end, the researcher examined components of both its program design and implementation as guided by the following three research questions:

1. Do USA TestPrep® benchmark assessments predict Oklahoma Core Curriculum Test performance?
2. Using Webb's Depth of Knowledge Framework, how well do the USA TestPrep® benchmark assessments align with the cognitive dimension of deeper learning?
3. How did teachers use and perceive the program to assist them in improving instruction and student learning?

An evaluation of these three research questions will help gain a more complete understanding of how well the assessment tool aligns with the policy goals of deeper learning and college and career readiness.

## Summary

The federal and state policy emphasis on producing college and career ready students has resulted in a flurry of local responses, such as the adoption of interim assessment systems, which attempt to align instruction and assessment with the knowledge and skills required for college and career readiness. Yet, many students remain ill-prepared for college and careers (Conley & Darling-Hammond, 2013; Bitter & Loney, 2015) and scholars continue to debate even how to define college and career readiness (Kobrin, 2007; Mayurama, 2012). While many states and districts are turning to interim assessment systems to help them facilitate instruction and monitor progress towards college and career readiness, there is little evidence to support the effectiveness of these programs in influencing student achievement (Perie, et al., 2009). The purpose of this study is to examine how well one interim assessment program, USA TestPrep®, is designed to meet the goals of deeper learning and college and career readiness through an examination of its alignment to deeper learning and predictive power as an interim assessment tool as well as its usefulness to teachers as a tool for instructional improvement.

The next section will synthesize assessment literature and advance the conceptual framework for the study, which includes a discussion of the underlying assumptions of deeper learning and how those assumptions will be used to analyze and interpret the study data. Next the report offers a description of the method used to answer the research questions; it also includes a discussion of the study sample, sources of data, and procedures for data collection, as well analytical approach. Findings resulting from each of the research questions are reported. Finally, the report then concludes with a discussion of the interpretation and significance of the findings, limitations of the study, and suggestions for future research.

# Chapter 2: Review of Literature

## Overview

Assessment is integral to teaching and learning. According to Wiliam (2011), learning outcomes often bear little resemblance to the intended objectives, and this highlights the critical role that assessment plays in effective instructional practice. Stiggins (2002) notes that assessment provides data that can illuminate the successes of students and teachers alike, as well as the system itself. Stiggins and Wiliams are not alone in their assertion that assessment is an important component in teaching and learning. Darling-Hammond et al. (2013) argue:

> Assessments can positively influence instruction through their diagnostic value, as well as by communicating important learning goals and modeling appropriate pedagogy. They can guide helpful interventions and teaching decisions. However, assessments can also have negative consequences if they are designed or used in ways that distort teaching, deny students access to learning opportunities from which they could benefit, or create incentives for schools to underserve or exclude students with particular needs. Thus, both the assessments themselves and the decisions related to their interpretation and use must be subject to scrutiny. (p. 13)

It is important that the decisions that policymakers and educators at all levels of the educational system make regarding assessment do not belie its importance. Accordingly, they must be judicious in assessment choices and continue to monitor and evaluate these choices via sound theory and empirical evidence.

Assessment initiatives are not limited to the state and federal level. Local policymakers also adopt assessment policy and programs, such as interim assessments, with the belief that these measures will increase student achievement and fulfill demands of state and federal mandates (Goertz, Oláh, & Riggan, 2009; Perie et al., 2009). Yet, increases in test scores can be due to familiarity with test content and not due to learning (Amrein & Berliner, 2002; Harlen, 2005; Berliner, 2011); once-a-year assessments do not close the achievement gap (Amrein and Berliner, 2002; Berliner, 2011; Stiggins, 2015). Furthermore, these assessment policies can often have detrimental side-effects. For example, a National Research Council (2011) study found that high-stakes testing policies can result in high test anxiety and low self-esteem, lower graduation rates and can distort achievement results due to altered curriculum and test preparation strategies - all without significantly improving student achievement.

There is evidence as to why these assessment and accountability mandates fail to result in the desired outcomes. Black and Wiliam (1998b) suggest that testing initiatives have failed to produce effective policy because they are not intended to provide direct support to classroom instruction. Noguera et al. (2015) found that not only do these initiatives fail to provide direct support to classroom instruction, they can result in "differential access" to curricula where minority and lower socio-economic status students are often placed in remedial courses, and contribute to failure to deliver the skills that students need to be prepared for college and career. Scholars also find further evidence for the failure of assessment policies in teachers' perceptions of assessment. Teachers tend to view assessment in terms of student accountability for desired outcomes rather than improving learning. Gulikers, Biemans, Wesselink and van der

Wel (2013) found that, “Teachers do not differentiate between formative and summative assessment; assessment is always seen as grading and certifying at the end of learning” (p. 122). Noguera (2015) argues that accountability distorts the use of testing from improving teaching and learning to monitoring and measuring achievement.

This study evaluates how a locally-adopted assessment system, USA TestPrep®, aligns with measures of deeper learning and supports instruction and facilitates further deeper learning activities both within the program and beyond. But before doing so, this section will provide an overview of assessment literature and its role in learning and instruction, focusing largely on formative assessment, summative assessment, balanced assessment, and interim assessment.

### *Formative and Summative Assessment*

Assessment is a judgment against a set of standards which is recorded in terms of a comparison or ranking (Scriven, 1967; Taras, 2005). Use of the term has evolved. For many years, the term assessment was used to describe a process that measures the ability of an instructional activity to produce the desired results (Wiliam, 2011). Those desired results are often based on a set expectation or standard. “A judgment cannot be made within a vacuum, therefore points of comparison, i.e. standards and goals, are necessary” (Taras, 2005, p. 467).

Assessments are typically divided into those which are formative and those which are summative in nature. In 1967, Scriven first coined the terms formative and summative evaluation when referring to the evaluation of education programs (Lau, 2016; Taras, 2005). Formative and summative assessment today encompasses much more than evaluation of programs. Application of the terms has evolved. Today, those terms most often are used in reference to measures of student achievement. The shift to this usage first occurred in 1971 when Benjamin Bloom introduced the terms formative and summative assessments when referencing student learning (Lau, 2016; Wiliam, 2011). Since Bloom first used formative and summative assessment to mark student achievement, scholars have studied and evaluated their role in teaching and learning.

## Formative Assessment

Black and Wiliam (1998a) argue that there is not a universally accepted term for formative assessment; scholars use terms such as classroom evaluation, classroom assessment, internal assessment, instructional assessment, and student assessment. Thus, Black and Wiliam (1998a; 1998b) distinguish formative assessment from assessment in general, by asserting that assessment refers to all the functions that teachers and students use to provide information to adapt teaching and learning and that it only becomes formative when the information is actually used to change learning activities to meet student needs. While educators may refer to many activities as formative assessment, many educators stop short of using the data to diagnose student learning and to adapt instruction in ways that meet students' learning needs (Ford, Van Sickle, & Fazio-Brunson, 2016; Marsh, Pane, & Hamilton, 2006; Stiggins, 2002).

Formative assessment can also refer to instruction and learning activities as well. Shepard (2005) emphasizes the importance of the instructional component of assessment when she asserts that formative assessment and scaffolding instruction are equivalent. The relationship between assessment and learning is evidenced by the terminology scholars use when discussing formative assessment. Linquanti (2014)

Stiggins (2002), Taras (2005),  (Wiliam, 2011) and Black, Harrison, Lee, Marshall and Wiliam (2004) all use the term assessment for learning when referring to formative instruction. Stiggins (2002), while acknowledging the synonymous use of formative assessment and assessment for learning, cautions that one should not do so. Assessment for learning does not merely provide data for educators; it includes students in the process by encouraging their sense of self-efficacy and in turn, their desire to continue their academic growth and development. Formative assessment and assessment for learning should be a cycle of collecting data from a number of sources and responding accordingly.

Additional scholarship has since built upon Black and Wiliam's (1998a) initial definitions of formative assessment, and this research continues to emphasize its holistic nature and its critical role in changing teaching and learning. Perie et al. (2009) contribute to this body of scholarship asserting that, "Formative assessment is used by classroom teachers to diagnose where students are in their learning, where gaps in knowledge and understanding exist, and how to help teachers and students improve learning" (p. 6). Formative assessment activities not only help to diagnose learning gaps but to identify where students are in relation to outcomes measured in summative assessments (Gulikers, Biemans, Wesselink, & van der Wel, 2013). It should be emphasized that the literature does not view formative assessment as a measurement tool in and of itself; but rather, it is a process that facilitates teaching and learning (Black & Wiliam, 1998a; Heritage, 2010; Linquanti, 2014). Shepard (2005) asserts that formative assessment is a collaborative transaction involving negotiation in order to improve learning outcomes; "Formative assessment is a dynamic process in which supportive adults or classmates help learners move from what they already know to what they are able to do next" (p. 66).

The discussion of formative assessment as a process shifts the attention from what formative assessment is to how it can be used and what it can help accomplish. Formative assessment can increase student achievement and result in significant learning gains (Black & Wiliam, 1998a). According to Fuchs and Fuchs (1986) in their meta-analysis of formative assessment when formative assessment practices were strengthened, not only did learning outcomes increase, but the effect size was larger than most other intervention strategies. Effective formative assessment strategies increase gains especially among low performing students and thus help to close the achievement gap (Black & Wiliams, 1998b, p. 141).

Yet, there is a *poverty of practice* in which classroom assessments are often rife with problems and fall short of intended outcomes (Black & Wiliams, 1998b). The literature addresses this widespread inability of the educational system to implement effective formative assessment practices. Some studies identify policy as the culprit in undermining effective formative assessment use. In these arguments, high-stakes testing and accountability have inhibited the productive use of formative assessment (Black et al., 2004) because such high emphasis has been on singular, summative assessments, which dominate teacher time and focus, and encourage teaching practice focused more on rote, low-level cognitive skills (Black & Wiliam, 1998b).

Others attribute this *poverty of practice* to lack of teacher capacity (Heritage, 2010; Stiggins, 2002). Stiggins (2002) goes as far to say that our nation's educators are "unschooled in the principals of sound assessment – be it assessment of or for learning" and it results in the misdiagnosis of student achievement and student learning needs (p. 762). Stiggins suggests an action plan that would include comprehensive professional development for both teachers and administrators; it would also add an assessment component to certification competencies and teacher and administrator preparation programs.

Heritage (2010) agrees that ineffective assessment practices partly stem from weak teacher capacity to use assessments for learning. She argues that focusing attention and resources on building teacher knowledge and skill rather than developing the best assessment tool would better serve educational achievement. Heritage is not alone in recognizing the link between instructional capacity, assessment practices, and student learning. Instruction and formative assessments are inseparable (Black & Wiliam, 1998b). Formative assessment is embedded in learning and is directly tied to current curriculum and teaching (Perie et al., 2009). Instructional capacity for skills such as incorporating student self-assessment and self-esteem (Black & Wiliam, 1998b), creating “a culture of questioning and deep thinking” (Black & Wiliam, 1998b), and effective feedback (Hattie & Timperley, 2007; Heritage, 2010; Knight, 2002; Stiggins, 2002) are crucial for effective formative assessment.

In summarizing formative assessment and its role in teaching and learning, there is a need for a shift in perspective and practice. Heritage (2010) argues, “Instead of considering formative assessment within the context of a measurement paradigm, perhaps we should be focusing on firmly situating the process of formative assessment within a learning paradigm” (p. 15). Formative assessment should focus on the process of learning and teaching and the skills that both teacher and student acquire in the process rather than measurement as a comparison or rank. This shift parallels the argument that Darling-Hammond and Adamson (2013) make for assessing for deeper learning. They argue that curricular and instructional systems that use assessments of deeper learning are essential; claiming, in fact, "assessment measures are designed to improve teaching and learning" (p. 14).

## Summative Assessment

Summative assessments fall on the opposite end of the continuum from formative assessment. These assessments have become entrenched in educational policy and practice due to the emphasis on high-stakes accountability. According to Black and Wiliam (1998b), they differ from the learning focus of formative assessment; summative tests spotlight overall summaries of student achievement rather than providing data to diagnose students' learning needs. As Linquanti (2014) notes, "Summative assessments render judgment after the conclusion of instruction, and can occur at the classroom or system level" (p. 6).

The utility of summative assessments is predicated at all levels of the educational system: national, state, district, or classroom. At the national level, summative assessments are defined based on national outcomes (Gulikers et al., 2013). They are used to measure student achievement across states for the purpose of informing policy (Perie et al., 2009; Stiggins, 2002). At the state level, Linquanti (2014) asserts that summative assessments focus on accountability in order to evaluate educational programs and/or measure student achievement; they are distant from the learning process and occur at a broader context. Summative assessments also occur at the district or classroom level. In this case, summative assessments may include end-of-the-unit or end-of-semester testing (Perie et al., 2009). They are intended to measure student progress or achievement and thus they focus on measurement of learning rather than assisting immediate learning needs (Linquanti, 2014). These tests are the least flexible of the assessment types and are generally used simply for grading purposes (Perie et al., 2009).

As discussed above, summative assessments may measure the effectiveness of educational programs or provide a large-scale view of student achievement at the national, state, district, and even building level. Heritage (2010) asserts that they are needed to in order to make valid and reliable measurements about how learners are doing in relation to standards. Stiggins (1999) acknowledges that summative assessments can incite students, teachers, and administrators to strive to meet high academic standards set forth in the standardized, high-stakes testing. Additionally, they can provide important data that are comparable across classrooms to those who make policy and program decisions.

Yet, the utility of summative assessments is not unquestionable, especially in a culture of high-stakes testing and accountability. Campbell (1976) in his study of social change implementation and measurement finds, “The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor” (p. 85). Nichols and Berliner (2008) apply this axiom, known as *Campbell’s Law*, to high-stakes testing in education. They assert that the intense pressure and consequences associated with high-stakes testing has corrupted the process it was intended to assess; this has resulted in unintended consequences such as narrowing of curriculum, teaching to the test, and cheating so that the results gained from test scores are a distorted representation of teaching and learning. Self-Determination Theory (SDT) literature suggests that this corruption occurs due to the controlling, rather than informational nature of the indicator (Adams et al., 2016; Ford, Van Sickle, & Fazio-Brunson, 2016; Ryan & Weinstein, 2009). Evidence of unintended consequences, distortions, and even abuses is found in the literature surrounding educational assessment policy related to high-stakes testing and accountability. Linn (2000) concludes that when accountability is attached to summative assessments, their utility is often outweighed by the unintended, and negative, consequences that arise.

One unintended consequence of summative assessments is that results may be distorted and not a true measure of student achievement. As referenced earlier in this chapter, Amrein and Berliner (2002) studied eighteen states with high-stakes testing policies to determine whether or not the accountability policy fostered transfer of learning. They found that while the state summative test scores may show increases, student achievement when compared to four standardized tests (ACT, SAT, NAEP, and AP tests) remains level or even actually decreases with the implementation of high-stakes testing. The rise in test scores was misleading and the intended outcome of improved student learning was not achieved. High-stakes testing has failed to yield the desired outcomes at an international level as well. The National Research Council [NRC] (2011) asserts that accountability policies have not helped to close the achievement gap between the United States and the highest performing countries.

The NRC (2011) in its synthesis of high-stakes testing policy offers insight into these results; concluding, “Incentives will often lead people to find ways to increase measured performance that do not also improve the desired outcomes" (p. 2). Harlen (2005) notes:

> This high-stakes use is universally found to be associated with teachers focusing on the content of the tests, administering repeated practice tests, training students in the answers to specific questions or types of question, and adopting transmission styles of teaching. In such circumstances, teachers make little use of assessment formatively to help the learning process. (p. 209)

The focus on high-stakes test preparation does yield improved results on summative tests themselves; however, this is not indicative of increased achievement. Amrein and Berliner (2002), Harlen (2005), and Noguera et al. (2015) attribute these results to a narrowed curriculum and test preparation.

Summative assessments can fall short in their measurement of learning in other ways. The NRC (2011) concludes these tests usually provide measures of performance only in tested subjects and grades, do not measure intangibles such as curiosity, persistence, and collaboration, nor measure distal goals of education, such as success in career, civic, or personal life (p. 37). Amrein and Berliner (2002) describe additional shortcomings of high-stakes summative assessment explaining that acquisition of knowledge is a proximal goal of education. The distal goal of education is transfer of learning to new and/or unique situation. This notion that application of knowledge to another context is more difficult to assess and is often far different from the outcomes measured in high-stakes testing (p. 13). Conley and Darling-Hammond (2013) echo this notion that current states' assessments are unable to measure deeper learning. The assessments often fail to provide useful information regarding the students' ability to think critically and are unable to produce useful data indicative of student preparation for college and career.

Not only do high-stakes summative assessments often fail to provide an accurate and complete measure of student learning, they can also produce results that are harmful to those whom they were intended to benefit. Amrein and Berliner (2002) argue that if, as their study demonstrates, high-stakes testing policies do not promote learning then they are futile to successful schools. Furthermore, the unintended consequences of a narrowed curriculum, increased drop-out rate, and higher retention rate, while not good for any student, are particularly harmful to economically disadvantaged and minority students (pp. 10 – 11). Noguera et al. (2015) also address the equity concerns that arise from high-stakes testing noting that disadvantaged students suffer the most in this process. They reason that:

> This has occurred because (1) in many schools, especially those serving low-income students, the curriculum has been narrowed to mirror the tests; and (2) test scores have been used to allocate differential access to the curriculum, with the result that students of color and low-income students have often been denied access to a thinking curriculum and instead relegated to remedial, rote-oriented, and often scripted courses of study. (p. 4)

The potential harm extends beyond academic inequity. Summative assessment can both directly and indirectly influence student motivation. Harlen (2005) finds that direct manifestations of high-stakes testing can be high levels of student test anxiety and low scores and this can influence self-esteem and student perceptions of themselves as learners. Betts and Costrell (2001) found that students with greater academic ability increased their efforts in response to high-stakes testing while students with less ability dropped out (as cited by NRC, 2011, p. 20). Indirectly, high-stakes testing influences teachers and curriculum (p. 210). Lazear (2006) notes, “As a policy issue, testing is as much about motivating teachers as it is about motivating students” (p. 1042). Lazear also finds that these high-stakes testing policies prompt some teachers to increase their effects while other educators respond by leaving the profession. Deci, Spiegel, Ryan, Koestner, & Kauffman (1982), in a study of teacher behavior, found that teachers responded to accountability by becoming more controlling and providing fewer opportunities for students’ autonomous learning, talked more - giving commands, criticizing and praising; all of which is detrimental to students’ intrinsic motivation.

In sum, summative assessments can be a useful tool to policy makers. They provide data that may allow for measurement of student achievement and progress and help evaluate program effectiveness (Heritage, 2010; Linquanti, 2014; Perie et al., 2009; Stiggins, 2002). However, the utility of summative assessments, especially in a high-stakes environment, can be marred by unintended consequences (Linn, 2010). Their ability to measure achievement may be compromised (Amrein & Berliner, 2002). For example, curriculum may be narrowed and unfavorable instructional practices such as rote learning and fewer opportunities for autonomous learning may result (Amrein & Berliner, 2002; Harlen, 2005; Noguera et al., 2015). They are limited in what they measure. They do not measure persistence, collaboration, or future success. Often, they fail to measure deeper learning and instead focus on lower cognitive skills (NRC, 2011). Both student and teacher intrinsic motivation can decrease as well (Deci et al., 1982).

## Interim Assessment

The challenges associated with formative and summative assessments need not thwart good teaching and learning. Scholars argue that assessments can be used in a smarter way in order to ensure that the desired goals are attained. While formative and summative assessments are often viewed as diametrically opposed, this notion is a false dichotomy. Lau (2016) echoes Taras (2005) when she too asserts that considering one form of assessment good and the other bad is erroneous. “While Scriven and Bloom both intended for summative and formative assessment (evaluation) to be linked and to work together, such a link was gradually lost” (Lau, 2016, p. 512).

However, recent literature has begun to once again recognize the important link between the two. A balanced assessment system that includes both summative and formative assessment is needed in order to guide students through the learning process (Guikers et al., 2013; Lau, 2016; Perie et al., 2009; Stiggins, 2002). Assessment influences instruction; teachers tend to instruct in the manner in which their students will be tested (Darling-Hammond & Adamson, 2013; Darling-Hammond & Conley, 2015). Stiggins (2002) argues that educators must learn to distinguish between assessment of learning and assessment for learning and understand that both are important to student achievement. Teachers need on-going support in how to interpret assessment data and then respond by adapting instruction (Faxon-Mills, Hamilton, Rudnick, & Stecher, 2013). Assessments are intended to improve teaching and learning; and therefore, they ought to build capacity.

To recapitulate, the description of assessment as a dichotomy of formative and summative assessment is erroneous (Lau, 2016). Rather, formative and summative assessments are interconnected. Biggs (1998) finds that students engage in the learning process when summative assessments and other learning opportunities, including formative assessment, are in line with one another (as cited in Lau, 2016, p. 518). While they each serve their own unique purposes, they also depend on one another to help propel student achievement (Gulikers et al., 2013; Lau, 2016; Perie et al., 2009; Stiggins, 2002).

Formative and summative assessments are not the only forms of assessment. States and local schools are increasingly utilizing interim assessment systems with the hope of improving student outcomes by evaluating and monitoring instructional programs and strategies and standardizing curriculum (Goertz et al., 2009; Perie et al., 2009). These assessments promise to help increase student achievement and fulfill the demands of policy (Goertz et al., 2009; Perie et al., 2009) by providing diagnostic information that, unlike high-stakes testing, allows districts and educators to modify instruction during the school year (Perie et al., 2009). On a continuum, summative being at the broad end and formative being at the narrow end, interim assessments are considered medium scale assessments (Perie, Marion, Gong, & Wurtzel, 2007). Interim assessments refer to tests that both evaluate student skills within a set time frame and against a set of standards and whose data can be aggregated at the school and district level (Geortz et al., 2009; Hamilton et al., 2009; Perie et al., 2009). Extending the continuum analogy, interim assessments are “middle-ground” assessments. They accomplish what separately formative and summative assessments cannot. Perie et al. (2007) elucidate the role of interim assessments declaring them “middle tier” assessments between formative and summative assessment. They have the advantage of providing prompt feedback to the classroom educator, as does formative assessment. But, unlike formative assessments, they can also provide aggregated data to inform district-level decisions much like summative assessments, but in a timely manner that allows schools to influence instruction (Perie et al., 2009).

As with formative and summative assessment, interim assessments have different names. Some include diagnostic, predictive, benchmark, and even formative assessment (Perie et al., 2009). Perie et al. (2007), in their policy brief on interim assessment, assert that the inability of many standardized summative assessments to influence instruction during the school year has prompted many districts to look to interim assessments to inform and audit student learning. In this brief, they identify three key purposes of interim assessments: predictive, instructional, and evaluative. With predictive objectives, interim assessment data serve to predict performance on future summative end-of-year assessments. When interim assessments have an instructional purpose, educators have the capacity to analyze and effectively use data to adapt teaching to meet student needs and enrich the curriculum. When the interim assessment has evaluative purposes, data are used to measure the success of programs, strategies, and teachers. The data do not inform immediate decisions; rather, they improve instructional programs over a period of time so that future students are the recipients of change.

As Perie et al. (2009) and Goertz et al. (2009) note, many schools operate under the belief that interim testing provides data that guide instructional practice and ultimately result in greater student achievement. Yet there is scant evidence documenting how schools actually use benchmark testing, how policies facilitate benchmark testing to improve learning, or the role of benchmark testing with other forms of assessment (Goertz et al., 2009). In fact, Shepard, Davidson, and Bowman (2011) studied mathematics teachers' use of interim testing data and found that teachers typically describe their use of data within a framework of accountability and not in terms of instruction and learning.

Nevertheless, the implementation of interim testing continues to rise. School districts are increasingly using interim assessments to increase achievement although its ability to do so is not empirically documented (Goertz et al., 2007; Heritage, 2010). In fact, Shepard et al. (2011), in their study for the National Center for Research on Evaluation, Standards, and Student Testing, found that interim tests did not measure higher levels of cognitive demand. Rather, the focus was on lower levels of cognitive skills. Darling-Hammond and Adamson (2013) and Conley and Darling-Hammond (2015) also note the deficit of critical thinking skills in assessment systems.

Interim assessments alone are too insufficient to guide instructional improvement; in order for interim assessments with an instructional intent to further student achievement, support and structures must be in place. Their use for improved teaching is aided by alignment with standards, district expectations that they would guide teaching, a useful information management system, time to reteach, and instructional support (Goertz et al., 2009). Interim assessments vary considerably and no single test can provide all the data educators need to make informed decisions; data systems should incorporate information from a variety of sources (Hamilton, Halverson, Jackson, Mandinach, & Supovitz, 2009).

As previously stated, interim assessment purposes must be clearly defined before a district can accurately determine the effectiveness of the assessment (Perie et al., 2009). Policymakers should be able to answer specific questions before adopting an interim assessment system. Doing so will help them develop a theory of action for how the assessment will improve student achievement. Perie et al. (2009) suggest the following questions are essential for delineating the purpose of interim assessments:

1) What do we want to learn from this assessment?
2) Who will use the information gathered from this assessment?
3) What action steps will be taken as a result of this assessment?
4) What professional development or support structures should be in place to ensure the actions steps are taken and are successful?
5) How will student learning improve as a result of using this interim assessment system and will it improve more than if the assessment system was not used? (p. 9)

In addition to a clear statement of purpose, district policy makers should consider a cost-benefit analysis to ensure that interim assessments provide information that is not otherwise readily available (Perie et al., 2007, p. 20). The consumption of precious district resources is not worth the cost if the interim assessment merely serves as a mini summative assessment and does not directly tie to specific instructional units and thus provide teachers with information (Perie et al., 2009). Identifying a clear purpose, developing a theory of action, and conducting a cost-benefit analysis will help ensure the quality of interim assessments and the decisions that result from the data generated. Finally, Perie et al. (2009) also suggest that future research on interim testing consider whether or not predictive assessments are accurate in their estimation of summative assessment performance. In other words: to what extent do assessments for instructional purposes improve instruction and thus measure student learning outcomes?

## The USA TestPrep® Program

In order to address some of the above identified gaps in the literature, such as the predictive power of interim assessments and their ability to improve instruction and student outcomes, the focus of this research study is on USA TestPrep®, an exemplar of one such district-adopted interim assessment system. USA TestPrep® is a standards-based online resource that is intended to help districts measure and improve student test performance. According to the program website, USA TestPrep®, a Georgia-based company, was founded in 1998 (“About Us,” 2018, para. 1) by two teachers who wanted to use technology to improve test scores (“Teacher Developed,” 2018, para. 1) and whose goal “is to allow students in the class to work on self-directed activities while other students can receive individual instruction from the teacher” (“Teacher Developed,” 2018, para. 2).

USA TestPrep® serves nearly 2 million student users and over 70,000 educators (“About Us,” 2018, para. 7) and can provide curriculum aligned to individual state standards as well as Common Core State Standards (“About Us,” 2018, para. 1). Among the program resources is the capability for standards-based benchmark testing. Additional program components include instructional videos, practice questions, review games and even printable worksheets. Students may track their own mastery and progress and work through the program components via self-directed activities. Teachers can track individual and whole-class progress and assign program components to individual students and whole classes. Both teachers and students may access the program at any time from home or school.

As an interim assessment system, USA TestPrep® can serve as a center-point for both formative and summative assessment. On the narrow end of the assessment continuum, the program provides for formative assessments; allowing individual classroom teachers to use data from benchmark testing and other instructional activities to adjust instruction based on individual student needs. But, towards the broader end of the assessment continuum, the program also allows data to be aggregated to the classroom, building, and even district level. As an interim assessment system, USA TestPrep® has the potential to address all three purposes of interim assessments. As stated, it may be used to inform instruction at the classroom, building, and district level. With an evaluative intent, the data gleaned from program components can be used to assess the success of teachers, strategies, and programs over a period of time. With a predictive purpose, benchmark test data may be used to predict scores on the state end-of-instruction assessment. This study's research questions consider each of these three purposes in the evaluation of USA TestPrep®.

## Summary

In sum, school districts are increasingly responding to accountability pressures by turning to interim testing (Goertz et al., 2009; Perie et al., 2009). Local policy makers operate under the belief that interim testing can provide data to increase student achievement (Goertz et al., 2009). Yet, there is a lack of empirical evidence to support this belief (Heritage, 2010). As Perie et al. (2009) propose, interim tests can serve three purposes: instructional, evaluative, and predictive. Districts should clearly identify the purpose of the benchmark so that they can determine the effectiveness of the assessment.

This section provided a brief history assessment and a synthesis of the empirical evidence regarding, formative, summative, and interim assessment. It also introduced the key components and theory of action for the interim assessment systems that are the focus of this study. The next section will provide a conceptual framework for how the district interim assessment policy and the program that executes it advance student achievement.

# Chapter 3: Conceptual Framework

## Overview

Educational systems seek to impart knowledge and prepare students for life beyond the classroom. At the federal, state, and local level, policies and programs are implemented with the intent of increasing student proficiency and preparing a workforce to drive the economy forward. Statutes such as the No Child Left Behind Act (2001) and now the Every Student Succeeds Act (2015) have resulted in policies including implementation of rigorous standards, high-stakes testing, and school accountability and more (Lee & Reeves, 2012; Linn et al., 2002).

Yet policies often fail to produce the desired results (NRC, 2011). Huberman, Bitter, Anthony, and O'Day (2014) note that No Child Left Behind's failure to realize the intended student outcomes has prompted much questioning and debate regarding what students must know and do in order to be productive citizens both at school and in the workforce. Students are leaving secondary schools ill-prepared to meet the demands that await them in tertiary education. American universities find entering college freshmen to be deficient in critical thinking and problem-solving skills (Conley & Darling-Hammond, 2013, p. 2). Policy makers and researchers continue to seek ways for schools to produce students who have both the content knowledge and skill set to be successful beyond the classroom. Increasingly these skill sets are identified and described as deeper learning (National Research Council [NRC], 2012; Huberman et al., 2014).

Deeper learning serves as the conceptual framework for measuring how USA TestPrep® not only measures student learning outcomes but also facilitates further opportunities for increased student achievement. Deeper learning's value as a framework, as will be more fully articulated in what follows, is based in its focus on higher-order cognitive skills and competencies that will transfer into college and career and other areas of productive citizenship.

The next section provides an overview of deeper learning. The discussion will begin with an explication of the domains and competencies of deeper learning and will continue with an examination of the underlying assumptions of deeper learning in instruction and assessment of student outcomes. It will conclude with a discussion on Webb's Depth of Knowledge as a framework by which the cognitive domain of deeper learning will be operationalized. Finally, this section will advance the study's research questions.

### *Deeper Learning Definition*

Deeper learning is a set of skills and competencies used to describe instruction, learning, and achievement. Twenty-first century skills, college and career readiness, student centered learning, next generation learning, new basic skills, higher order thinking, and meaningful learning are terms associated with this way of describing desired school processes and outcomes (NRC, 2012, p.1). Deeper learning is the ability to master content-specific knowledge and then use that knowledge in novel situations (Bitter & Loney, 2015; Huberman et al., 2014; NRC, 2012). Much of the literature references the National Research Council's definition of deeper learning:

> We define "deeper learning" as the process through which an individual becomes capable of taking what was learned in one situation and applying to new situations (i.e., transfer). …Through deeper learning (which often involves

> shared learning and interactions with others in a community), the individual develops expertise in a particular domain of knowledge and/or performance. The product of deeper learning is transferable knowledge including content knowledge in a domain and knowledge of how, why, and when to apply this knowledge to answer and solve problems. (NRC, 2012, pp. 5–6)

In a synthesis of the literature on deeper learning, the NRC (2012) found that the characteristics of deeper learning could be classified into three competency domains: cognitive, interpersonal and intrapersonal (see Table 1). The William and Flora Hewlett Foundation (2013) asserts that deeper learning focuses on the interaction of six competencies nested within the three domains: mastery of core academic content, critical thinking and complex-solving skills, effective communication skills, collaboration skills, an understanding of how to learn, and academic mindsets.

The cognitive domain focuses on the types of knowledge and how it is organized in an individual mind (NRC, 2012). Deep content knowledge as well as critical thinking and complex problem solving are two dimensions associated with the cognitive domain (William and Flora Hewlett Foundation, 2013; Huberman et al., 2014). Additional proficiencies comprise the cognitive domain. Those include cognitive process and strategies, knowledge and creativity (NRC, 2012).

Two clusters of competencies structure the interpersonal domain: collaboration and leadership (NRC, 2012). The sociocultural perspective argues that participation in society influences the individual's learning. Skills from the interpersonal domain are reflective of the sociocultural perspective that holds that knowledge is acquired within a social context. Competencies such as teamwork, collaboration, empathy, self-

**Table 1. Deeper Learning Defined**

| Domain | Clusters | Student Proficiencies |
|---|---|---|
| Cognitive | Deep content knowledge<br>Critical thinking and complex problem-solving | Procedural knowledge of content area<br>Apply core knowledge to new tasks<br>Formulate and solve problems<br>Data analysis and statistical reasoning<br>Creativity and non-linear thinking |
| Interpersonal | Collaboration<br>Leadership | Teamwork<br>Collaboration<br>Empathy<br>Self-presentation<br>Social influence |
| Intrapersonal | Learning to learn<br>Academic mindsets | Flexibility<br>Cultural appreciation<br>Social responsibility<br>Initiative<br>Perseverance<br>Self-regulation |

*Note:* Adapted from NRC (2012) and William and Flora Hewlett Foundation (2013).

presentation, and social influence all contribute to deeper learning in the interpersonal domain. Huberman, Bitter, Anthony, and O'Day (2014) echo this thought when it identifies two of its six dimensions of deeper learning as belonging to the interpersonal domain: collaboration and communication.

While the interpersonal competency acknowledges that participation in a culture may influence deeper learning, the individual himself also influences deeper learning. Dweck and Legget assert that student beliefs about learning may strongly influence learning outcomes (as cited NRC, 2012). This concept forms the foundation of the intrapersonal domain. The NRC (2012) proposes that the intellectual openness, work ethic, and conscientiousness clusters organize the intrapersonal domain and that deeper learning is aligned with the intrapersonal traits of intellectual openness, work ethic, and core self-evaluation. More specifically, skills such as flexibility, cultural appreciation, social responsibility, initiative, perseverance, and self-regulation all play a role in mastery of content and deeper learning. Two competencies associated with deeper learning come from the intrapersonal domain: learning-to-learn and academic mindsets (Huberman et al., 2014).

For purposes of this study, this discussion on the nature of deeper learning will focus on the cognitive domain. As discussed in the previous chapter, interim assessments are becoming an increasingly frequent response to the pressures to improve student outcomes, despite the insufficiency of data to support their ability to do so (Goertz et al., 2007; Heritage, 2010). The first step in assessing program efficacy is determine if it measures what it purports to measure: the predictive power of the interim assessment system. The ability of USA TestPrep® to measure and improve student test performance is based upon student proficiencies that fall within the cognitive domain of deeper learning. With a focus on college and career readiness, policy makers and researchers strive to develop those outcomes in students who not only have content knowledge but are prepared to use that deep content knowledge to find success beyond the classroom. Those student outcomes are now more than ever being measured in terms of not only academic content knowledge but the ability to think critically and problem-solve (NRC, 2012; Huberman et al., 2014). While the interpersonal and intrapersonal domains of deeper learning are important, a focus on the cognitive domain allows the researcher to consider how interim assessments improve student learning outcomes as measured by academic content knowledge, critical-thinking, and problem-solving skills and in turn contribute to developing students who are college and career ready.

## Cognitive Competencies

While there is a paucity of research regarding the concept of deeper learning itself (NRC, 2012), there does exist a body of research regarding the competencies and components that embody deeper learning. It is important to note that cognitive components of deeper learning are not novel ideas. They are aspects of human competence that have been valued for centuries (NRC, 2012).

As previously stated, a key competency of deeper learning is transfer. Transfer is commonly defined as the use of previously acquired information in a new and novel ways (Bitter & Loney, 2015; Huberman et al., 2014; NRC, 2012). It is the outcome that results from deeper learning (NRC, 2012). The Gesalt psychologists studied this concept of transfer. They differentiated between rote learning and meaningful learning and found evidence for a relationship between meaningful learning and transfer (NRC, 2012). According to Katona (1942), meaningful learning, or understanding, promotes transfer whereas rote learning, or memorization, does not. Instruction based on memorization is not successful. Rather, students should be engaged in content and continually use information in complex ways that scaffolds learning so that they may generate new knowledge in unique situations (William and Flora Hewlett Foundation, 2013). Simple recall and recognition tasks do not promote the more complex processing that is necessary for transfer to occur (NRC, 2012).

Cognitive competencies are not fixed traits; they are malleable and change over time (NRC, 2012). In order to acquire these characteristics, students must be provided opportunities to develop these traits and receive appropriate feedback for deeper learning to occur (Bitter, Taylor, Zeiser, & Rickles, 2014; NRC, 2012). Feedback is information provided by an agent such as a teacher, peer, or even experience regarding aspects of one's performance or understanding (Hattie & Temperly, 2007). The frequency and nature of feedback with respect to creating opportunities for deeper learning is critical. The practice must be pervasive and feedback must permit students to self-correct (NRC, 2012). Hattie finds that feedback must help students develop metacognition and generate ambitious goals (as cited in Fullan, 2015, p. 277).

While the nature of deeper learning promotes transfer and is influenced by feedback, the cognitive domain competencies are shaped by specific knowledge and skill sets. These skills must scaffold and be interactive for transfer and deeper learning to occur (NRC, 2012). Deeper learning in the cognitive domain occurs when learners master core academic content knowledge and think critically (William and Flora

Hewlett Foundation, 2013). Huberman et al. (2014) expand upon the description of deeper learning:

> Mastering core academic content means that students have learned and can recall relevant facts from a content area; have procedural knowledge of content area; can use the language specific to a content area; and can apply core knowledge to new tasks and situations in other academic subjects, to real-world situations, and in non-routine ways. (p. 9)

Bitter and Loney (2015) note that mastering core academic content signifies that there is a baseline level of knowledge from which students must create and produce transferable knowledge. In addition, that procedural knowledge allows students to value the problem-solving skills that facilitate that transfer of knowledge to novel situations.

Critical thinking skills is another competency within the cognitive domain. The ability to think critically and solve complex problems occurs when, "Students apply tools and techniques gleaned from core subjects to formulate and solve problems. These tools include data analysis, statistical reasoning, and scientific inquiry as well as creativity, nonlinear thinking, and persistence" (William and Flora Hewlett Foundation, 2013, p. 2). These problem-solving skills require students to amalgamate knowledge from a variety of resources (Bitter & Loney, 2015).

The American Institutes for Research (AIR) identified four characteristics of schools that were committed to deeper learning. The AIR study selected eleven pairs of schools in California and New York. Schools that were identified as highly experienced in and committed to deeper learning, referred to as networked schools, were matched with schools with similar demographics that did not identify as committed to and experienced in deeper learning. In an analysis of the structures and cultures of schools that facilitate deeper learning, the study found that four goals were a key component in creating opportunities for mastering core academic content and developing critical thinking skills: explicit goals for developing cognitive competencies, curriculum drawn from a specific set of standards, instruction that incorporates real-world situations, and long-term assessments (Huberman et al., 2014).

A second, follow-up report considered the extent to which students experienced opportunities for deeper learning. The proposed theory of action suggests that schools with a culture and structure that support deeper learning create more opportunities for students to engage in deeper learning. The authors' analysis confirmed this (Bitter et al., 2014). The remainder of this chapter will discuss what opportunities for deeper learning can be found in instruction and assessment and concludes by advancing a framework for identifying those measures of deeper learning in the cognitive domain.

## Deeper Learning in Interim Assessment

After a description of deeper learning and its composite skills and competencies, the focus now shifts to how these assumptions play out in the theory and practice of student learning and the role of interim assessment systems in this process. DuFour and DuFour (2015) note:

> The United States must recognize that if students are to learn at deeper levels, schools must create the conditions that allow for on-going, deeper learning of the educators who serve those students in each of the three critical areas – (1) curriculum, (2) pedagogy, and (3) authentic assessment. (pp. 30-31)

Schools must be cognizant of the conditions necessary for deeper learning and then provide the support and structure to foster those conditions.

The role of deeper learning in current assessment systems is important for a number of reasons. As previously discussed, current policies fail to produce the hoped-for improvements in student learning. The consideration of appropriate assessment systems is a key part of a well-balanced policy initiative; but, as some scholars note, the current assessment systems of many states are largely ineffective in identifying and describing student college and career readiness (Darling-Hammond & Adamson, 2013; Darling-Hammond & Conley, 2015).

The efficacy of assessment systems to measure deeper learning is not a concern in and of itself; assessment influences instruction, and this too shapes quality student learning. Teachers tend to instruct in the manner in which their students will be tested (Darling-Hammond & Adamson, 2013; Herman & Linn, 2014). Schools ought to be knowledgeable about the relationship between assessment systems and instruction. Access to assessment data alone does not stimulate effective use of data to modify instruction. Teachers need on-going support in how to interpret and then respond by adapting instruction (Faxon-Mills et al., 2013). The relationship between assessment and instruction is not tenuous; it is substantial. As stated in the review of literature, Darling-Hammond and Adamson (2013) assert the imperative for educational systems that use assessments of deeper learning. Assessments are intended to improve teaching and learning and, therefore, they ought to build capacity.

The American Institutes for Research, via funding from William and Flora Hewlett Foundation, studied deeper learning in schools. This study, mentioned previously, compared nineteen “networked” schools that identified themselves as committed to and experienced in deeper learning with twelve similar “non-networked” schools mainly in California and New York. The researchers found that schools identified as committed to and experienced in deeper learning reported more frequent use of formative assessments as well as more traditional summative assessments. Project-based learning, portfolios, exhibitions, collaborative, long-term assessments and student defense of artifacts were some of the strategies used to develop deeper learning in the cognitive domain (Bitter & Loney; 2015; Huberman et al., 2014).

With revenue and funding shortfalls, districts are going to have to rely more and more on existing resources. They must consider how well existing assessment measures meet instruction, learning, and assessment needs. High quality assessments can be affordable and feasible and given the financial conditions of many schools in the United States, it is more imperative that assessment is cost-effective (Darling-Hammond & Adamson, 2013). As discussed in the previous chapter, Goertz et al. (2009) argue for more extensive research that studies how interim assessments are implemented and considers the quality of data generated by those assessments. Given the constraints on the fiscal and other resources of schools and the demand for more research into the efficacy of interim assessments in producing student outcomes, this study sets out to consider the utility of a specific interim assessment system to produce the desired student outcomes and meet instruction, learning, and assessment needs as identified by the cognitive dimension of deeper learning.

Part of the empirical learning process included taking the elements of deeper learning, specifically those related to the cognitive domain, and overlaying them with the benchmark tests as provided for by USA TestPrep® to examine how well the benchmark tests align with deeper learning. An analysis of the benchmark tests measure of deeper learning provided a crucial link to how educators might use data gleaned from the program to facilitate further opportunities for deeper learning both within the confines of the USA TestPrep® program and beyond. The next section will discuss Webb's Depth-of-Knowledge as a framework to examine how well USA TestPrep® benchmark tests align with components of deeper learning.

## Webb's Depth of Knowledge and Cognitive Demand of Assessments

Cognitive hierarchy schemes have been the basis for the development of and assessment of curriculum and standards for some time (Webb, 2010). Norman Webb traced a brief history of cognitive hierarchies. His analysis finds that from earlier schemes such as the work of Ralph Tyler's process for analyzing curriculum, Bloom's Taxonomy of intellectual behaviors, and Loren Anderson's revised taxonomy to more recent work such as a 2007 TIMMS study and a 2009 PISA all share a strong foundation in expressing content complexity.

Webb's Depth-of-Knowledge [DOK] is the framework by which the cognitive domain of deeper learning will be identified and described in the course of this study. Webb's DOK is the most conventional and established method of establishing the cognitive requirements of test items (Yuan & Le, 2014). Wise and Alt (2006), Herman, Webb, & Zuniga (2007), Herman & Linn (2013), Yuan & Le (2014), and Herman, La Torre Matundola, & Wang (2015) have all used Webb's DOK as the method for measuring cognitive demand.

The Depth-of-Knowledge scale was developed to elucidate content complexity (Webb, 2010). It refers to the cognitive demands and complexity required by an item (Herman et al., 2007; Webb, 1997) and was specifically designed in order to evaluate the relationship between assessments and expectations (Webb, 1999; 2007; 2010). Cognitive complexity of an item should be distinguished from item difficulty:

> Difficulty in assessment is a statistical term related to the proportion of students who answer an assessment task correctly. Difficulty can be associated with other factors such as exposure to instruction, opportunity to learn, and other than home language that are not related to content complexity, a characteristic that is more associated with the content structure" (Webb, 2010, p. 17)

Students may experience difficulty in answering assessment questions even though the item itself is associated with a low level of cognitive complexity (Herman et al., 2007; Wyse & Viger, 2011).

Content complexity is characterized by several key traits. It is a continuum that is based on content analysis rather than cognitive analysis (Webb, 2010). Webb's DOK levels content complexity into four categories (Webb, 2007):

- Level 1 (recall) includes recalling information such as a fact, definition, term, or a simple procedure as well as performing a simple algorithm or applying a formula.
- Level 2 (skill/concept) includes the engagement of some mental processing beyond a habitual response. A level 2 assessment item requires students to make some decisions as how to approach the problem or activity.

- Level 3 (strategic thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at level 3. Activities that require students to make conjectures are also at this level.
- Level 4 (extended thinking) requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. (pp. 11-12)

Webb's DOK does provide more precise classifications for each of the four content areas (Webb, 1997, 2002, 2010).

It is also important to consider what DOK is not. Depth of Knowledge does not depend on verbiage that can be misleading and lead to inaccurate classifications (Webb, 2010). DOK does not fundamentally alter if the population changes (Wyse & Viger, 2011). It does not stipulate a way in which students may respond to assessment items. They may use a variety of strategies to generate a response (Wyse & Viger, 2011).

## Summary

Assessment systems are a form of capacity tool. Capacity tools provide education, information, and resources to the target audience (Schneider & Ingram, 1990). Stakeholders at all levels of the educational system from policy makers to educators and parents and students expect to gain knowledge and feedback from assessment results (Herman et al., 2007).

To be useful, the information needs to be perceived as relevant and useful to the end-user—in this case, the teacher. Assessments of deeper learning are critical to our students and nation (Darling-Hammond & Adamson, 2013). Furthermore, to reap the benefits of a high-quality assessment system, the state and local components of an assessment system must work together to develop an assessment system that yields instructionally useful information (Darling-Hammond & Adamson, 2013). The researcher seeks to determine to what extent USA TestPrep®, a local component of the assessment system in a single district, yields instructionally useful information regarding the deeper learning of students in its district. To arrive at the program's ability to yield instructionally useful information, the researcher first examined the purpose of the interim assessment system and its ability to measure what it purports to measure: academic content knowledge.

But, scholars assert that assessment systems must provide more than just a measure of academic content knowledge; they be measures deeper learning and provide data that influence teaching and learning (Darling-Hammond & Adamson, 2013; Darling-Hammond, Wilhoit, & Pittenger, 2014). Thus, the researcher extended the study to consider the efficacy of the interim assessment system to not only fulfill its purpose and measure academic content knowledge but provide relevant data concerning deeper learning in students and finally how teachers can use that assessment system to guide learning and instruction in developing students who are college and career ready.

# Chapter 4: Method

The purpose of this study was to examine features of the design and implementation of the USA TestPrep® system as an exemplar of a typical interim assessment program and to gather empirical evidence of their utility in helping districts, schools, and teachers facilitate deeper learning experiences and prepare students to meet the demands of college and careers. The following questions guided the study:

1. Do USA TestPrep® benchmark assessments predict Oklahoma Core Curriculum Test performance?
2. Using Webb's Depth of Knowledge Framework, how well do the USA TestPrep® benchmark assessments align with the cognitive dimension of deeper learning?
3. How did teachers use and perceive the program to assist them in improving instruction and student learning?

## Research Design

This study, by virtue of the focal research questions, was designed as a mixed-methods study. The mixed method approach is "a research paradigm whose time has come" (Johnson & Onwuegbuzie, 2004, p. 14) and combines both quantitative and qualitative sources of data. Firestone (1987) asserts quantitative and qualitative approaches, while distinct from one another, can complement each other. The complementarity of a mixed method approach adds additional strength to a study above and beyond a singular qualitative or quantitative approach (Creswell & Plano Clark, 2007; Johnson & Onwuegbuzie, 2004; Yin, 2006). Johnson and Onwuegbuzie (2004) posit that mixed methods research accomplishes this by providing pictures and narrative that enhance the meaning of numbers while using the numbers to define and clarify the pictures and narrative. Thus, the complementarity of findings provides stronger evidence of the phenomenon under study that can illuminate conclusions that might be missed by a more one-dimensional approach.

**Table 2. Overview of Research Design**

| | Research Question | Analytical Approach | Data Sources |
|---|---|---|---|
| Research Question 1 | Do the USA TestPrep® benchmark assessments predict Oklahoma Core Curriculum Test performance? | Quantitative: Bivariate correlation | Oklahoma Core Curriculum Test results; USA TestPrep® data base |
| Research Question 2 | Using Webb's Depth of Knowledge Framework, how well do the USA TestPrep® benchmark assessments align with the cognitive dimension of deeper learning? | Qualitative: Content analysis | USA TestPrep® benchmark test questions |
| Research Question 3 | How did teachers use and perceive the program to assist them in improving instruction and student learning? | Quantitative: Descriptive Statistics<br>Qualitative: Analysis of Themes | Quantitative: Teacher survey responses;<br>Qualitative: Focus groups; Teacher survey responses |

Table 2 outlines the linkages between the study research questions and the data sources and analytical techniques used to answer those questions. The components of each will be more explicitly detailed in the following sections of this chapter. In order to answer research question one and evaluate the programs ability to yield instructionally useful information, the researcher studied USA TestPrep®'s ability to meet its purpose in predicting OCCT performance and in measuring what it purports to measure: academic content knowledge. The researcher employed bivariate correlational analysis to examine the relationship between two continuous variables: USA TestPrep® benchmark assessments and OCCT achievement test scores.

Assessments systems must provide measures of deeper learning (Darling-Hammond & Adamson, 2013; Darling-Hammond, Wilhoit, & Pittenger, 2014). Norman Webb's Depth of Knowledge is the framework used to explore the extent to which deeper learning is assessed with USA TestPrep®. The researcher explored this question using content analysis to calculate the total number and percent of assessment items for each level of depth of knowledge on each benchmark test.

Finally, assessment systems are a capacity tool and must not only provide data, the data must be relevant and used to change teaching and learning (Schneider & Ingram, 1990; Herman et al., 2007; Darling-Hammond & Adamson, 2013). The researcher employed both quantitative and qualitative approaches to consider research question three: How did teachers use and perceive the program to assist them in improving instruction and student learning? For the quantitative component, the researcher examined descriptive statistics of responses to a teacher survey that was conducted using Qualtrics, an online survey platform. In regard to the qualitative component, the researcher included free response portions in the survey that were designed to gain insight into use and perceptions of USA TestPrep®. The researcher conducted a thematic analysis of the qualitative data drawn from the free-response portion of the survey.

Additionally, the researcher examined an analysis of themes from focus groups comprised of teachers who participated in the USA TestPrep® program. The researcher designed focus groups to elicit teacher teachers' beliefs and perceptions of USA TestPrep® as a data driven decision-making (DDDM) tool to guide instruction. The protocol for focus groups can be seen in Appendix B. Taken together, evidence gathered from this mixed methods approach helped provide a more complete understanding of USA TestPrep®'s effectiveness and utility.

### Study Sample and Units of Analysis

The population for this study was all middle school teachers and students from a suburban district in northeastern Oklahoma. For research question one, this included approximately 850 students in grades seven and eight for each of the three years included in the study. Data sources came from the district's USA TestPrep® benchmark testing results for the school years 2013-2014, 2014-2015 and 2015-2016. Seventh grade students completed benchmarks in three subjects: geography, reading, and math. For purposes of this study, data analysis included only reading and math scores. Each subject was tested three times each school year resulting in 2,952 math scores and 3,556 reading scores.

Eighth grade students completed tests in four subjects: US history, reading, math, and science. As with seventh grade, this study included only eighth grade reading and math scores. Each subject was tested three times per school year resulting in 2,081 eighth grade math scores and 3,682 eighth grade reading scores. The researcher also used additional scores from the district's 2014, 2015, and 2016 seventh and eighth grade reading and math OCCT tests: 1,041 seventh grade math, 1,255 seventh grade reading, 761 eighth grade math, and 1276 eighth grade reading results.

The unit of analysis for research question two was question items from benchmark assessments created using the USA TestPrep® program. Assessments included as part of the study were seventh grade on-level math, seventh grade honors math, eighth grade math, seventh grade reading, and eighth grade reading given during the 2013-2014, 2014-2015 and 2015-2016 school years. The district gave benchmark exams three times per school year resulting in forty-four benchmark tests. The researcher could not locate one benchmark exam for seventh grade on-level math given during the 2013-2014 school year and thus did not include it in this study.

As mentioned previously, research question three concerned teachers' use and perceptions of the program to assist them in improving instruction and student learning. Thus, the researcher included teachers who used USA TestPrep® as part of their instructional duties during the 2013-2014, 2014-2015 and 2015-2016 school years as the sample and unit of analysis. Other school officials such as administrators and instructional coaches who used the program and teachers who did not use the program were not invited to participate.

Thirty-three teachers used the program during these years and were eligible for participation. At the time of the study, thirteen teachers who used the program during these years were no longer employed in the district. The researcher was able to contact three of those and invite them to participate in the study. Twenty teachers who used the program were still employed within the district and were available to participate in the study. A total of twenty-three teachers were available for participation in the study. Thirteen of the available twenty-three teachers elected to complete survey responses, resulting in a 57% participation rate. Four teachers indicated that they would like to participate in the focus group discussion. Three of those teachers were present for the focus group discussion, resulting in a 13% participation rate.

According to Peak and Fothergill (2009), group size is the key to focus group success and groups with three to five members tend to run better than larger groups. Krueger and Casey (2000) assert that while ideally a focus group should be comprised of six to eight participants, smaller groups are better suited for gaining in-depth insight and understanding. As stated earlier, the purpose of research question three was to gain a comprehensive understanding of how teachers used and perceived USA TestPrep® as a tool for data driven decision making. While the study focused on a three-year period, many of the teachers used the program for longer than three years and had a great deal of experience to share..

## Data Collection Procedures

The researcher gained school district permission to access all needed data sources and conduct the study. Additionally, the researcher secured Institutional Review Board consent from the University of Oklahoma.

The researcher collected seventh and eighth grade math and reading benchmark test results from the USA TestPrep® database during the school years 2013-2014, 2014-2015, and 2015-2016. Additionally, the researcher collected seventh and eighth grade math and reading OCCT results for 2014, 2015 and 2016. The researcher entered student USA TestPrep® scores and their corresponding OCCT scores for each of the three academic years of the study into an SPSS database. The researcher conducted a bivariate Pearson's correlation analysis including a measure of statistical significance.

For research question two, the researcher collected each of the benchmark tests for seventh and eighth grade math and reading for the academic years 2013-2014, 2014-2015, and 2015-2016. The program design requires each assessment item be identified by depth of knowledge level. The research identified the total number of assessment items at each level of depth of knowledge for each benchmark assessment included in the study. The researcher then used these data to calculate what percent of each benchmark assessment corresponded to each level of depth of knowledge.

For research question three, the researcher sent recruitment emails to eligible participants that included a link to a Qualtrics survey regarding use and perceptions of USA TestPrep® as a tool to assist in improving instruction and student learning. After an initial email inviting those teachers to participate, the researcher sent four follow-up reminders over the course of two months before the survey was closed. Teacher participation was voluntary. Participants gave electronic consent at the beginning of the survey prior to viewing the first question. Survey items were developed to explore teacher perceived purposes of the interim assessments and the program; was the program perceived to serve instructional, evaluative, or predictive purposes? Related to perceived purposes, survey components were developed to measure which program components were used and how frequently. Finally, survey items were developed to exlore teacher perceptions of the program's ability to not only measure the cognitive dimension of deeper learning but also to provide opportunities for the teachers to use program components and data to help further develop the cognitive dimension. A copy of the survey and its contents is provided in Appendix A. The researcher used Qualtrics reporting to measure descriptive statistics for survey response items. The researcher also conducted an analysis of themes from the free-response portion of the survey.

As part of the teacher recruitment email sent to teachers who used USA TestPrep® during the years included in the study, the researcher also included an invitation to participate in the focus group. Interested teachers were invited to contact the researcher for further information. Any teacher who used the program at the study site during the years included in the study was invited to participate. Participation in teacher focus groups, as with the survey, was voluntary. The focus group was conducted off campus and at a neutral site to maintain the privacy and confidentiality of participants. Signed consent was given prior to beginning the focus group discussion.

Krueger and Casey (2000) describe the role of questions in driving the focus group conversation. Opening questions start the conversation and do not elicit profound information. Two to five key questions drive a focus group and ending questions facilitate reflection on the conversation and allow participants to comment on what they find to be of importance. With this in mind, the researcher designed questions that would begin a conversation on USA TestPrep®, then focus on teacher beliefs and attempt to gain information regarding program usage: why do (or why don't) teachers find the program to be a valuable tool, and finally, prompt participants to reflect on the program and contribute additional information that they felt to be of importance. The focus group protocol is provided in Appendix B. The researcher used an audio recorder to record the focus group conversation. Additionally, the researcher took hand-written notes as the participants shared their thoughts.  Immediately following the focus group, the researcher took notes regarding personal reflections of the focus group discussion. According to Miles, Huberman, and Saldaña (2014) raw data such as audio recordings must be transcribed into text in order to be analyzed. Accordingly, later that day, the researcher transcribed the focus group conversation into a word document. Krueger and Casey (2000) note that transcripts of focus groups, along with field notes taken by the researcher, are the basis for focus group analysis. Using the unabridged transcription of the focus group discussion as well as the hand-written notes taken during the focus group, the researcher began the analysis. The specific techniques used will be further explicated in the next section on analytical approaches.

## Analytical Approaches

The researcher took a mixed-methods approach using variety of analytical techniques. Research question one examines the predictive validity of the USA TestPrep® interim assessments. Predictive validity is a type of criterion-related validity that quantifies a correlation coefficient with a future criterion (Vogt, 2007, p. 120). This measure of correlation does not indicate what drives the relationship between the variables; it only describes the nature of the relationship. In other words, correlation analysis measures the symmetric relationship of two variables and does not indicate a causal direction (Vogt, 2007, p. 151). In this study, the two variables are scores on benchmark tests and OCCT results. In approaching research question one, do the USA TestPrep® benchmark assessments predict OCCT test performance, the researcher used SPSS to examine the correlation between two variables: USA TestPrep® benchmark assessment results and Oklahoma Core Curriculum Tests (OCCT) achievement test scores. The researcher conducted a Pearson's Correlation analysis using IBM SPSS

Statistics software to determine the nature and extent of the linear relationship between benchmark assessment scores and Oklahoma Performance Index (OPI) scores on the 7th grade math and reading and 8th grade math and reading OCCT assessments for each year of the study. The null hypothesis for the Pearson Product moment correlational analysis is $\rho = 0$. Table 3 provides a brief overview of 7th and 8th grade OCCT math and reading achievement descriptive statistics for the 2014, 2015, and 2016 school years.

**Table 3. OCCT Math and Reading Test Score Descriptive Statistics by Grade and Year**

| | 2014 | 2015 | 2016 |
|---|---|---|---|
| 7th Grade Math | 758.42 $\bar{x}$, 79.963 *s*<br>range: 400-990 | 758.86 $\bar{x}$, 72.395 *s*<br>range: 400-990 | 757.86 $\bar{x}$, 71.653 *s*<br>range: 514-990 |
| 7th Grade Reading | 762.79 $\bar{x}$, 69.725 *s*<br>range: 547-990 | 759.84 $\bar{x}$, 61.451 *s*<br>range: 556-924 | 763.81 $\bar{x}$, 71.800 *s*<br>range: 438-990 |
| 8th Grade Math | 741.66 $\bar{x}$, 62.457 *s*<br>range: 458-876 | 740.03 $\bar{x}$, 70.951 *s*<br>range: 400-990 | 736 $\bar{x}$, 71.240 *s*<br>range: 400-867 |
| 8th Grade Reading | 731.67 $\bar{x}$, 77.517 *s*<br>range: 494-990 | 786.96 $\bar{x}$, 70.427 *s*<br>range: 547-990 | 787.84 $\bar{x}$, 83.295 *s*<br>range: 489-990 |

The researcher used content analysis in order to analyze the results of research question two: Using Webb's Depth of Knowledge Framework, how well do the USA TestPrep® benchmark assessments align with the cognitive dimension of deeper learning? Do they measure critical thinking skills and deep content knowledge and, if so, to what extent? A measure is considered valid if it measures what it purports to measure (Kelley, 1927; Warner, 2013). Content validity measures the extent to which the content of a test or survey aligns with the content it is intended to measure (Vogt, 2007, p. 118). It addresses whether or not items in a test or survey represent all theoretical dimensions or content areas; whereas the more subjective face validity is concerned with if the items *appear* to measure what they say they measure (Warner, 2013, p. 939). Webb's Depth of Knowledge will be used as a framework in order to determine how well the content of benchmark tests align with measures of deeper learning.

Webb's criteria for alignment are based upon four measures: categorical concurrence between standards and assessment items, range of knowledge, balance of representation between objectives on standards and assessment items, and depth of knowledge (Webb, 2007). Webb (2007) utilizes four categories for interpreting depth of knowledge: level 1 (recall), level 2 (skill/concept), level 3 (strategic thinking) and level 4 (extended thinking). In this way, the study will identify the depth of knowledge in assessment items to determine how well they align with measures of deeper learning.

The researcher used both quantitative and qualitative data to answer research question three: How did teachers use and perceive the program to assist them in improving instruction and student learning? In the quantitative portion of the analysis, the researcher conducted a frequency analysis of survey responses collected in Qualtrics, an online survey platform. The researcher also conducted an analysis of themes from two free response questions on the survey and from a focus group comprised of teachers who used USA TestPrep® as part of their instructional duties during the years 2013 – 2014, 2014 – 2015 and 2015 – 2016. The researcher, as Kreuger and Casey (2000) suggest, began the analytical process by reading the transcript of the focus group to in order to bring to mind once again the entirety of the free response items and focus group discussion.

Prior to conducting the analysis, the researcher generated a list of codes to identify themes and the context in which they were used. The survey free response items focused on benchmark creation and use.  Codes generated for benchmark use responses included “Instructional,” “Evaluative,” and “Predictive.” The researcher coded responses based on using benchmarks for instructional intent such as identifying weaknesses and strengths or content and students in need of remediation. The researcher used “evaluative” to code responses relating to teacher, strategy, and/or program effectiveness. The researcher coded responses relating to predicting OCCT performance as “Predictive.” Codes related to responses describing how benchmark tests were created include “mimic” for program-generated tests that mimic OCCT in breadth and depth and “custom” for responses indicating that teachers created each test by hand-selecting questions. Miles et al. (2014) describe subcoding as codes that enhance and refine primary codes. The researcher used “DOK” and “OBJ” as sub-codes to identify responses related to depth of knowledge and standards and objectives. Focus group codes included “Testing,” “Resources,” “Process,” and “Data,” to categorize responses related to using USA TestPrep® testing components, other program resources, the benchmark testing process, and data generation and use. The researcher used the subcodes “Positive” and “Negative” to categorize responses that indicated how the participant felt about each of the primary codes.

The researcher then began the process of identifying and analyzing themes. Concerning the identification of themes, the researcher used the navigation pane in the word document to conduct a word count of unique words that were used by participants in their responses. Krueger and Casey (2000) assert that frequency does not necessarily equate importance and infrequency does not signify that insight is not important. According to Ryan and Bernard (2003), identifying repetitions is a simple way to find themes but the researcher must decide at what point a repetition becomes an important theme. Miles, Huberman, and Saldaña (2014) find that counting not only helps the researcher quickly identify information and verify hunches, it also helps keep qualitative research unbiased. Ryan and Bernard (2003) point out that identifying the context in which each repetition is used and sorting those context into similar meanings helps keep those repetitions in the context in which they were used. Thus, the researcher then studied the context in which the word was used each time and identified those contexts according to similar meaning.

## Data Triangulation

Triangulation, or using multiple methods and data sources, allows researchers to be more confident of their results and produce more valid findings (Mathison, 1988). The mixed methods approach of this study is one way to help provide a more complete picture of the study and its findings. Additionally, the design of data sources for research question three with both quantitative and qualitative components helped to triangulate data.

One way in which the researcher triangulated data was through survey response items. The researcher included both quantitative and qualitative response items that addressed how program components were used. For example, one item was a free-response question that asked teachers to describe how they created benchmark assessments. Were the assessments program generated tests that mimic the OCCT or were they created by teachers who selected topics and questions to include?

Additionally, the researcher included quantitative response items that addressed which program components were used and how often, the perceived purpose of benchmark assessments, and program ability to measure deeper learning in students. These items were designed to contribute to the findings and discussion in research questions one and two regarding the relationship between benchmark assessments and OCCT performance and the alignment of those benchmark assessments with the cognitive dimension of deeper learning.

The design of the focus group was another way in which data triangulation was achieved. According to Miles et al. (2014), the goal should be to use sources of triangulation that have different foci and strengths. Qualitative data from a focus group was collected in order to either complement and hopefully corroborate the quantitative data collected in the survey items or draw out differences for future analysis. The survey items and focus group protocol were designed in such a way that would help highlight the similarities and differences with one another as well as illuminate the findings in research questions one and two.

## Role of the Researcher

The researcher was an employee of the district in which the study took place and may have worked with potential participants as a teaching colleague and/or instructional coach. Furthermore, as an instructional coach for the local district, the researcher was responsible for improving instruction. In that capacity, had access to OCCT results and was responsible for training in and management of the USA TestPrep® program. The researcher did not teach in a content area that used USA TestPrep® to benchmark student achievement or participate in OCCT testing. Miles et al. (2014) assert that characteristics of a good qualitative researcher include familiarity with the phenomenon and setting under which the study takes place (p. 42). In this sense, the researcher's familiarity with the setting and program aided the researcher as an instrument in the qualitative portion of the study.

Personal biases have the potential to influence the course of the study. Threats to objectivity might include a number of ethical issues. The researcher recognized and remained aware of personal biases and avoided their influence by not only acknowledging their existence but by implementing a well-designed study, using appropriate sampling techniques that employ a variety of participants, subgroups, and data to help maintain objectivity and accuracy of results. Additionally, the advice of experienced researchers such as the dissertation chair and committee helped to avoid bias and conduct ethically-sound research.

# Chapter 5: Findings

The purpose of this research was to examine the utility of USA TestPrep®, an interim assessment system, in extending teacher capacity in meeting the learning needs of students and preparing them to be college and career ready. Findings of this study allow the researcher to examine components of both program design and implementation and to consider how they facilitate delivering deeper learning to students. These three questions guided the study:

1. Do the USA TestPrep® benchmark assessments predict Oklahoma Core Curriculum test performance?
2. Using Webb's Depth of Knowledge Framework, how well do the USA TestPrep® benchmark assessments align with the cognitive dimension of deeper learning?
3. How did teachers use and perceive the program to assist them in improving instruction and student learning?

The findings from the analysis of study data described in the previous chapter are presented in this chapter and are organized by research question.

## Research Question 1

This question addressed the degree to which the USA TestPrep® benchmark exams predict Oklahoma Core Curriculum Test (OCCT) performance. In order to consider the predictive validity of the benchmark assessments, the researcher collected 7th and 8th grade math and reading benchmark assessments created on USA TestPrep® during the 2013 – 2014, 2014 – 2015, and 2015 – 2016 school years. The researcher then collected 7th and 8th grade math and reading and entered the OCCT results along with their corresponding benchmark assessment results into IBM SPSS Statistics software. Both the benchmark and OCCT results are continuous variables, with any number of possible values. As such, the researcher ran a Pearson's correlational analysis to examine the relationship between two continuous variables: benchmark assessment scores and OCCT outcomes.

Pearson's correlation coefficient describes the relationship between two variables. Coefficients range on a scale from -1 to +1. A value of 0 indicates that there is no relationship between the two variables. Negative values indicate that as one variable increases, the value of the other variable decreases. A positive Pearson's *r* value indicates that as variable *X* increases so does variable *Y*. The closer the value is to an absolute score of 1, the more closely one can predict the *Y* value from the *X* value (Warner, 2013, p. 264). In the case of this study, the closer the *r* values are to an absolute score of 1, the more closely benchmark assessment scores predict OCCT performance.

While Pearson's correlation measures symmetry between two variables, as stated earlier, this measure of symmetry does not indicate a causal relationship. Correlational analyses measure the association between two variables and not a causal direction (Vogt, 2007). The results of the analysis of research question one outcomes are illustrated in Table 4 and Table 5, and a detailed description of these results follows.

**Table 4. Pearson's Correlation Analysis of OCCT Math Scores and USA TestPrep® Interim Assessments**

| 7th Grade On-Level Math | | | |
|---|---|---|---|
| | 2013 – 2014 | 2014 – 2015 | 2015 - 2016 |
| Benchmark 1 | *r* = .641 | *r* = .668 | *r* = .426 |
| Benchmark 2 | *r* = .656 | *r* = .688 | *r* = .723 |
| Benchmark 3 | *r* = .680 | *r* = .643 | *r* = .700 |
| **7th Grade Honors Math** | | | |
| | 2013 – 2014 | 2014 – 2015 | 2015 – 2016 |
| Benchmark 1 | *r* = .376 | *r* = .408 | *r* = .637 |
| Benchmark 2 | *r* = .580 | *r* = .505 | *r* = .495 |
| Benchmark 3 | *r* = .568 | *r* = .600 | *r* = .569 |
| **8th Grade Math** | | | |
| | 2013 – 2014 | 2014 – 2015 | 2015 – 2016 |
| Benchmark 1 | *r* = .527 | *r* = .570 | *r* = .574 |
| Benchmark 2 | *r* = .634 | *r* = .624 | *r* = .525 |
| Benchmark 3 | *r* = .613 | *r* = .705 | *r* = .582 |

*Note.* Two-tailed significance of *p* = .000 was achieved for all *r* values included in the table.

### *Mathematics*

Based on findings of a correlational analysis, seventh grade on-level math benchmarks demonstrated a stronger positive correlation with seventh grade math OCCT performance than does seventh grade honors math. Pearson correlation coefficient values for seventh grade on-level math range from *r* = .426, p < .001 to *r* = .723, p < .001. Eight of the nine benchmarks had a Pearson correlation coefficient of *r* = .641 or higher which would be indicative of a strong positive relationship. Of the three math courses, 7th grade on-level math demonstrates consistently stronger *r* values for all benchmark tests and OCCT performance.

Seventh grade honors math benchmarks are positively related to seventh grade math OCCT performance. Pearson correlation coefficient values range from $r = .376$, p $<$ .001 to $r = .637$ p $<$ .001. Six of the nine benchmarks had a Pearson correlation coefficient of $r = .505$ or higher which would be indicative of a moderate to strong positive relationship. Furthermore, eighth grade math benchmarks have a moderate to strong positive correlation with 8th grade math OCCT performance. Pearson's correlation coefficients ranged from $r = .525$, $p < .001$ to $r = .705$, $p < .001$. While the maximum $r$ value was not as high as seventh grade on-level math, eighth grade math demonstrates a higher minimum $r$ value.

**Table 5. Pearson's Correlation Analysis of OCCT Reading Scores and USA TestPrep® Interim Assessments**

| | 7th Grade Reading | | |
|---|---|---|---|
| | 2013 – 2014 | 2014 – 2015 | 2015 – 2016 |
| Benchmark 1 | $r$ =.674 | $r$ =.668 | $r$ =.699 |
| Benchmark 2 | $r$ =.658 | $r$ =.698 | $r$ =.692 |
| Benchmark 3 | $r$ =.606 | $r$ =.702 | $r$ =.681 |
| | 8th Grade Reading | | |
| | 2013 – 2014 | 2014 – 2015 | 2015 – 2016 |
| Benchmark 1 | $r$ =.635 | $r$ =.614 | $r$ =.656 |
| Benchmark 2 | $r$ =.525 | $r$ =.536 | $r$ =.641 |
| Benchmark 3 | $r$ =.636 | $r$ =.570 | $r$ =.610 |

*Note.* Two-tailed significance of p = .000 was achieved for all $r$ values included in the table.

### *Reading*

An analysis of seventh grade reading benchmarks and OCCT performance findings result in a range of Pearson correlation coefficients from $r = .606$, p $<$ .001 to $r$

= .702, $p < .001$. This is indicative of a strong positive relationship between benchmark assessment and OCCT performance. The correlation coefficients for eighth grade reading benchmarks and OCCT performance, as with reading, indicate a positive relationship. Pearson correlation coefficients ranged from $r = .525$, $p < .001$ to $r = .656$, $p < .001$. While this is indicative of a strong positive correlation, the correlations are not as strong as those for seventh grade reading benchmarks and OCCT performance.

### *Summary*

All benchmark assessments included in this study demonstrated a positive relationship with OCCT performance. However, when taking into account the $r^2$ value (coefficient of determination) and the wide variation in correlation coefficients, the results are not as robust as might be expected. Math benchmarks demonstrated a $r^2$ value of .142 to .523. Ten benchmarks had a $r^2$ value ≥ .400, nine had a $r^2$ value ≥ .300, two demonstrated a $r^2$ value ≥ .200 and three $r^2$ values were $r^2$ value ≥ .100. Reading benchmarks demonstrated a $r^2$ range of .275 - .493. Twelve benchmarks demonstrated a $r^2$ value ≥ .400, four had a $r^2$ value ≥ .300 and two had a $r^2$ value ≥ .200. The coefficients at the high end of the range explain approximately 50% of the variance in OCCT and benchmark performance. The vast majority of results explain approximately 40% of the variance and at the low end of the range only 14% of the variance in test performance. The implications of these results will be addressed in the discussion.

Additionally, correlation coefficients did not demonstrate any patterns over time, between benchmarks, or across subject areas/courses. Correlation coefficients increased and decreased from benchmark to benchmark and year-to-year among all five courses included in the study. As the academic year progresses and instructional time increases, one might expect that the relationship between benchmark and OCCT performance would strengthen. This is not the case. A discussion of this significance will take place the subsequent chapter on discussions.

### Research Question 2

To answer research question two, the researcher examined how well the USA TestPrep® benchmark assessments align with the cognitive dimension of deeper learning. USA TestPrep® program design is such that benchmark test questions are correlated to and identified by the depth of knowledge intended by the Department of Education for students to be college and career ready (Comprehensive Solution, 2017). Data collection involved two steps. First, the researcher referenced the actual benchmark assessments created using the USA TestPrep® program and calculated the number questions of each benchmark test that corresponded to each depth of knowledge level. Once the number of questions at each DOK had been tallied, the researcher then calculated what percent of the benchmark test items were dedicated to each level of depth of knowledge.

Step two involved comparing the percent of each level of depth of knowledge for each benchmark assessment with the percentages on the corresponding OCCT according to the specified blueprints identified in the test and item specification. In table 6, the researcher reported the depth of knowledge blueprint as indicated in Test and Item Specifications. The researcher illustrated this comparison for 7th grade and 8th grade OCCT math and reading depth of knowledge with each of the course related benchmarks for school years 2013 – 2014, 2014 – 2015 and 2015 – 2016.

**Table 6. Test and Item Specifications Depth of Knowledge Range for OCCT Exams**

| | DOK 1 | DOK 2 | DOK 3 |
|---|---|---|---|
| 7th Grade Math | 10 – 15% of OCCT | 65 – 70 % of OCCT | 15 – 25% of OCCT |
| 7th Grade Reading | 10 – 15% of OCCT | 65 – 70 % of OCCT | 15 – 25% of OCCT |
| 8th Grade Math | 10 – 15% of OCCT | 65 – 70 % of OCCT | 15 – 25% of OCCT |
| 8th Grade Reading | 10 – 15% of OCCT | 65 – 70 % of OCCT | 15 – 25% of OCCT |

The researcher reported the number of question items and total percent at each level of depth of knowledge for each benchmark included in the study. The researcher included math results in Table 7 and reading results in Table 8.

### *Mathematics*

Three math course benchmark exams were considered as part of this research, seventh grade on-level math, seventh grade honors math, and eighth grade math. According to Test and Item Specifications, the depth of knowledge distribution for both seventh and eighth grade math OCCT is as follows: depth of knowledge level one is 10 - 15% of the test, depth of knowledge level two is 65 – 75% of the test and depth of knowledge level three is 15 – 25% of the test.

Seventh grade on-level benchmarks fell within the provided distribution for depth of knowledge level one in two of the eight available benchmarks. They exceeded the range of distribution for depth of knowledge level one in six of the eight available benchmarks. The distribution for depth of knowledge level two fell below the distribution range on five of the eight available benchmarks and exceeded the distribution range in three of the eight available benchmarks. All eight the benchmark assessments fell below the distribution range for depth of knowledge level three, with five of the eight assessments not having a single question at depth of knowledge level three.

## Table 7. Depth of Knowledge on Math USA TestPrep® Interim Assessments

| | | | | |
|---|---|---|---|---|
| | 7th Grade On-Level Math | | | |
| | **Domain (Target)** | **DOK 1 (10-15%)** | **DOK 2 (65-70%)** | **DOK 3 (15-25%)** |
| 2013 - 2014 | Benchmark 1 | 9Q – 45% | 11Q – 55% | 0Q – 0% |
| | Benchmark 2[a] | | | |
| | Benchmark 3 | 12Q – 31% | 25Q – 64% | 2Q – 5% |
| 2014 - 2015 | Benchmark 1[b] | 5Q – 17% | 14Q – 71% | 0Q – 0% |
| | Benchmark 2[b] | 5Q – 19% | 16Q – 59% | 0Q – 0% |
| | Benchmark 3 | 6Q – 15% | 33Q – 83% | 1Q – 2% |
| 2015 - 2016 | Benchmark 1[b] | 5Q – 16.5% | 14Q - 47% | 0Q – 0% |
| | Benchmark 2 | 5Q – 22% | 18Q – 78% | 0Q – 0% |
| | Benchmark 3 | 4Q – 12% | 29Q – 85% | 1Q – 3% |
| | 7th Grade Honors Math | | | |
| 2013 - 2014 | Benchmark 1 | 4Q – 16% | 20Q – 80% | 1Q – 4% |
| | Benchmark 2 | 8Q – 35% | 15Q – 65% | 0Q – 0% |
| | Benchmark 3 | 10Q – 25% | 26Q – 65% | 4Q – 10% |
| 2014 - 2015 | Benchmark 1 | 4Q – 17% | 19Q – 79% | 1Q – 4% |
| | Benchmark 2 | 3Q – 12.5% | 21Q – 87.5% | 0Q – 0% |
| | Benchmark 3 | 5Q – 13% | 35Q – 87% | 0Q – 0% |
| 2015 - 2016 | Benchmark 1 | 5Q – 20% | 19Q – 76% | 1Q – 4% |
| | Benchmark 2 | 0Q – 0% | 23Q – 100% | 0Q – 0% |
| | Benchmark 3 | 3Q – 10% | 30Q – 90% | 0Q – 0% |
| | 8th Grade Math | | | |
| 2013 - 2014 | Benchmark 1 | 4Q – 16% | 21Q – 84% | 0Q – 0% |
| | Benchmark 2 | 7Q – 26% | 20Q – 74% | 0Q – 0% |
| | Benchmark 3 | 7Q – 21% | 26Q – 79% | 0Q – 0% |
| 2014 - 2015 | Benchmark 1[b] | 3Q – 50% | 9Q – 35% | 0Q – 0% |
| | Benchmark 2 | 7Q – 26% | 20Q – 74% | 0Q – 0% |
| | Benchmark 3 | 3Q – 9% | 32Q – 91% | 0Q – 0% |
| 2015 - 2016 | Benchmark 1 | 4Q – 16% | 21Q – 84% | 0Q – 0% |
| | Benchmark 2 | 5Q – 25% | 15Q – 75% | 0Q – 0% |
| | Benchmark 3 | 3Q – 10% | 27Q – 90% | 0Q – 0% |

*Note.* Q = number of test questions at each level of depth of knowledge. [a]Benchmark was unavailable for the study. [b]Benchmark included teacher created responses that were not identified by depth of knowledge.

Findings were similar for seventh grade honors math. Three of nine benchmark assessments fell within the distribution range for depth of knowledge level one. One assessment did not have any question items from level one and five exceeded the distribution range for depth of knowledge level one. In regard to depth of knowledge level two, two benchmark exams fell within the distribution range. The remaining seven assessments exceeded the distribution range. None of the benchmark exams fall within the distribution range for depth of knowledge level three. Four exams fell below the distribution range and the remaining five exams did not have a question item at level three depth of knowledge.

In eighth grade math, only one assessment fell within the distribution, and at that only on one level of depth of knowledge. One benchmark assessment fell within range for depth of knowledge level one, seven exceed the range, and one fell below the distribution range. In regard to depth of knowledge level two, one benchmark exam fell under the distribution range and eight assessments exceeded the distribution range. No question items were at the depth of knowledge level three in any of the nine eighth grade math benchmark exams.

### *Reading*

In seventh grade reading, three of the benchmark assessments fell within the distribution range for depth of knowledge level one. Five assessments exceeded the range and one fell below the distribution range. All nine of the benchmark tests exceeded the distribution range for depth of knowledge level two. Eight assessments fell below the distribution range for depth of knowledge level three, with one of those eight not having any question items that were identified as level three. One benchmark assessment fell within the distribution range for depth of knowledge level three.

Findings were similar for eighth grade. Five of the nine eighth grade reading benchmark exams fell within the distribution range for depth of knowledge level one. The remaining four assessments just exceeded the distribution range. All nine of the benchmark exams exceeded the distribution range for depth of knowledge level two and fell below the distribution range for depth of knowledge level three. One exam did not have any question items from level three.

**Table 8. Depth of Knowledge on Reading USA TestPrep® Interim Assessments**

| | | | | |
|---|---|---|---|---|
| | | 7th Grade Reading | | |
| | Domain (Target) | DOK 1 (10-15%) | DOK 2 (65-70%) | DOK 3 (15-25%) |
| 2013 - 2014 | Benchmark 1 | 3Q – 12% | 21Q – 84% | 1Q – 4% |
| | Benchmark 2 | 5Q – 20% | 19Q – 76% | 1Q – 4% |
| | Benchmark 3 | 4Q – 16% | 18Q – 72% | 3Q – 12% |
| 2014 - 2015 | Benchmark 1 | 3Q – 12% | 21Q – 84% | 1Q – 4% |
| | Benchmark 2 | 5Q – 20% | 19Q – 76% | 1Q – 4% |
| | Benchmark 3 | 4Q – 16% | 18Q – 72% | 3Q – 12% |
| 2015 - 2016 | Benchmark 1 | 4Q – 16% | 19Q – 76% | 2Q – 8% |
| | Benchmark 2 | 3Q – 12% | 22Q – 88% | 0Q – 0% |
| | Benchmark 3 | 1Q – 4% | 20Q – 80% | 4Q – 16% |
| | | 8th Grade Reading | | |
| 2013 - 2014 | Benchmark 1 | 4Q – 17% | 18Q – 75% | 2Q – 8% |
| | Benchmark 2 | 3Q – 12% | 19Q – 76% | 3Q – 12% |
| | Benchmark 3 | 2Q – 11% | 17Q – 89% | 0Q – 0% |
| 2014 - 2015 | Benchmark 1 | 4Q – 17% | 18Q – 75% | 2Q – 8% |
| | Benchmark 2 | 3Q – 12% | 19Q – 76% | 3Q – 12% |
| | Benchmark 3 | 3Q – 12% | 19Q – 76% | 3Q – 12% |
| 2015 - 2016 | Benchmark 1 | 4Q – 17% | 18Q – 75% | 2Q – 8% |
| | Benchmark 2 | 3Q – 12% | 19Q – 76% | 3Q – 12% |
| | Benchmark 3 | 4Q – 16% | 19Q – 76% | 2Q – 8% |

*Note.* Q = number of test questions at each level of depth of knowledge.

### *Summary*

Reading benchmark assessments reflected DOK distributions that most closely matched the OCCT blueprints. They consistently included question items at levels one, two and three of depth of knowledge. Math benchmarks included significantly fewer level three questions than reading benchmarks, often failing to include question items from that level. Although, the majority of the benchmarks for both subjects fell below the distribution range for level three. The study included 44 math benchmark assessments with a total of 706 assessment items. A mere two percent of those items measured depth of knowledge level three and none of those assessment items measured depth of knowledge level four.

In the majority of benchmark assessments, both content areas failed to fall within the target distribution range for all three levels of depth of knowledge as set forth in the Test and Item Specifications. Yet, while the range distributions were off, both reading and math included the majority of questions from level two of depth of knowledge distribution. This is consistent with the Test and Item Specifications for OCCT reading and math. The significance of these findings will be discussed in the following chapter.

## **Research Question 3**

The researcher studied teachers' use and perception of USA TestPrep® as a tool to aid in instruction and student learning. Two data sources were used in the analysis: teacher survey responses and teacher focus groups. The researcher measured descriptive statistics of the quantitative component responses and conducted an analysis of themes for the qualitative component of the survey as well as for the focus group responses.

### *Analysis of Survey Data*

Thirteen teachers completed the survey. One portion of the survey was designed to determine teacher use of the USA TestPrep® program and its various components. Respondents answered questions regarding how frequently they used program components. This data is included in Table 9. The researcher found that teachers most consistently agreed that USA TestPrep® was used as a benchmark assessment tool. Eleven of thirteen teachers reported using USA TestPrep® for benchmark assessments one to three times per semester. The second highest consensus was in using student data reports. Six of thirteen teachers reported using USA TestPrep® to generate student data reports one to three times per semester. In regard to other program components, three or fewer teachers reported using any one component on a daily, weekly, or even monthly basis. In fact, the majority of teachers reported never using numerous program components.

Qualitative data from the survey corroborated these quantitative results. One free response item asked teachers which program resource(s) they used most often and for what purpose. Two participants did not respond to this item. One respondent did not specify a program component, responding only with “reinforcement and review.” The remaining ten of the thirteen responses mentioned the assessment component of the program.

**Table 9. Frequency Analysis for USA TestPep® Program Components**

| | Daily | Weekly | Monthly | 1 - 3 Per Semester | Never |
|---|---|---|---|---|---|
| Bell Ringer[a] | 0 | 3 | 0 | 1 | 8 |
| Benchmark Tests[a] | 0 | 1 | 0 | 11 | 0 |
| Free Response Items[a] | 0 | 1 | 2 | 0 | 7 |
| Games | 1 | 2 | 3 | 3 | 4 |
| High Scoreboard[a] | 1 | 0 | 0 | 0 | 9 |
| Item of the Day[a] | 0 | 2 | 0 | 0 | 9 |
| Performance Task Questions[a] | 0 | 0 | 1 | 3 | 6 |
| Projector Questions[a] | 0 | 0 | 1 | 3 | 6 |
| Standards-based Formative Assessments[a] | 0 | 0 | 2 | 4 | 4 |
| Instructional Videos[a] | 0 | 0 | 2 | 0 | 9 |
| Student Data Reports[a] | 0 | 1 | 1 | 6 | 3 |

*Note.* Some survey items had no response.

Four of those ten responses did not indicate for what purpose they used the assessment component. Three indicated that they used it because it was mandated to do so. One response indicated that the assessment component was used for test prep purposes, one indicated benchmarks were used to assess English Language Arts skills, and another response indicated that the depth of knowledge levels on questions were better on USA TestPrep® than in the textbook. A second free response component asked teachers to identify which program component they used the least and why. Once again, two participants chose not to respond to this item. Eight participants did not mention a specific program component. Three respondents mentioned that they used games the least, with one of those three also including the instructional videos and assignments in his/her response. The most common reason given for not using the program was lack of time or resources. Six participants indicated that they did not have time to use the program. Two participants did not give a reason for why they used the indicated program the least. One response mentioned lack of access to computer labs and unfamiliarity with the program other than using it for benchmark purposes. One mentioned the difficulty in modifying for his/her class and another indicated that the standards on USA TestPep® did not align to the state standards. Overall, the results from these two items indicate that teachers did not use or rarely used a majority of the program components. Teachers most frequently and commonly used the benchmark assessment tool.

The researcher also included survey items designed to elicit teachers' perceptions of USA TestPrep® as a tool to drive DDDM. Teachers responded to these questions using a four-point Likert-type scale ranging from strongly disagree to strongly agree. A score of one or two indicates disagreement and scores of three or four indicate agreement. The researcher designed these questions to describe the perceived purpose, support, utility, and outcomes of USA TestPrep®.

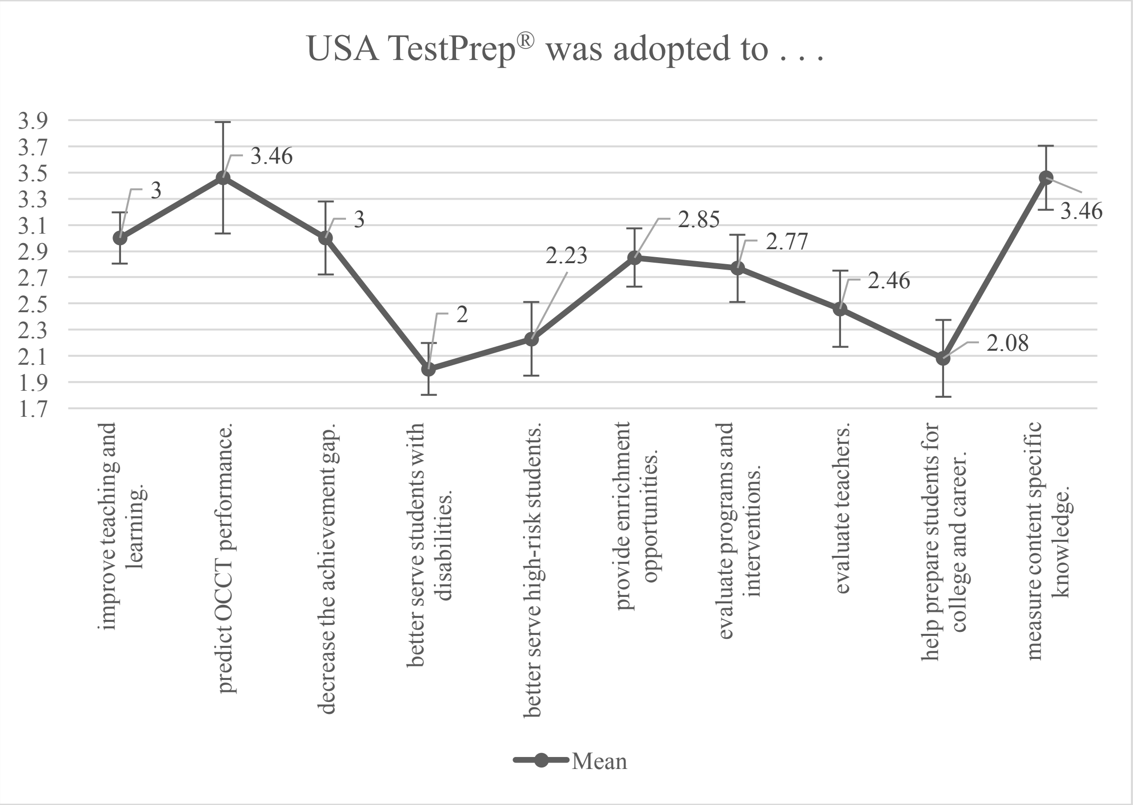


***Figure 1*. Descriptive Statistics for the Purpose of USA TestPrep®. The bars above and below the mean depict a 95% confidence interval for the estimate of the mean.**

An analysis of these data revealed that a majority of teachers, 92.3%, believe that USA TestPrep® was adopted to improve learning and also to predict OCCT performance. However, there was much less consensus regarding whether or not USA TestPrep® was adopted to close the achievement gap (69.3%), better serve students with disabilities (23.1%) or high-risk students (46.2%). Seventy-seven percent of teachers believe that USA TestPrep® was adopted to provide enrichment opportunities. Seventy-seven percent of teachers also believe that USA TestPrep® was adopted to determine if particular programs, curricula, or interventions are making a difference. Roughly half of the teachers surveyed believe that USA TestPrep® was adopted to evaluate teachers. Nearly all the teachers, 92.3%, believed that USA TestPrep® was adopted to measure subject-specific knowledge while only 38.5% of participants agreed that USA

TestPrep® was adopted to prepare students for college or career. In Figure 1, the researcher illustrated the mean score and confidence intervals for perceived purposes of the program. Confidence intervals allow for an estimate of the population parameter (Warner, 2013). Overlap in the intervals suggests that they are not statistically different. No overlap indicates that the categories are likely statistically different.

Figure 1 illustrates several points of consideration. Both predicting OCCT performance and measuring content-specific knowledge have similar intent in perceived purpose for USA TestPrep®. They both have a mean score of 3.46 and have confidence intervals that fall within the range for agreement. The confidence interval is noticeably tighter for the perceived purpose that USA TestPrep® was adopted to measure content-specific knowledge. In fact, it falls completely within the parameters for predicting the OCCT performance. The overlap indicates that the two values are not statistically different. The wide range of the confidence level for predicting OCCT performance is likely due to the small sample size.

The mean score for the intended purpose of decreasing the achievement gap was three with the possible score range being one to four. However, in regard to serving populations that would help close the achievement gap, there appears to be a disconnect. The mean score and confidence intervals for serving students with disabilities, serving high-risk students, and providing enrichment opportunities have not only lower means, but their confidence intervals are completely out of the range of closing the achievement gap, indicating a likely statistical difference. Finally, teachers indicated that they do not agree that USA TestPrep® was adopted to help prepare students who are college and career ready. In fact, the mean score is just above the strongly disagree range and both the upper and lower limits of the confidence interval are within the range of disagreement. In fact, the confidence levels, along with those of serving students with disabilities and serving high-risk students demonstrate no overlap with improving teaching and learning and indicate a likely statistical difference.

Survey prompts were designed to elicit teachers' perceptions of support provided for implementation and use of USA TestPrep®. While 85% of teachers believe that school administration had a clear plan for USA TestPrep® and encouraged its use, that number fell to 46% of teachers who believe that school administration modeled the use of data. Sixty-nine percent of teachers believed that school administration provided resources and/or support for use of USA TestPrep®. In regard to specific support and resources for use of USA TestPrep®, 46% of teachers reported receiving initial training in the use of USA TestPrep®, 54% reported being provided with on-going support in the use of USA TestPrep®, and 69% of reported receiving support in interpreting data from USA TestPrep® and using that data to inform teaching and learning in the classroom. Descriptive statistics for these prompts are illustrated in Figures 2 and 3.

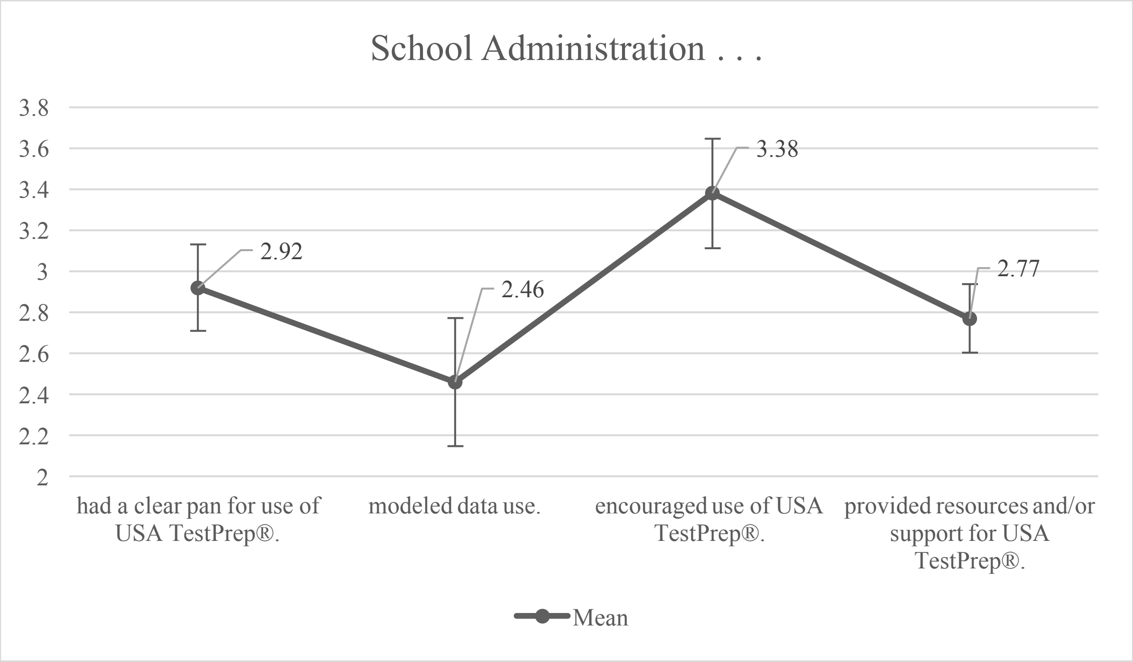


***Figure 2*. Descriptive Statistics for Administrative Support for USA TestPrep®. The bars above and below the mean depict a 95% confidence interval for the estimate of the mean.**

Of particular importance to the confidence intervals depicted in Figure 2 is the mean and confidence intervals both fall within the range for agreement that teachers that perceived the school supported the use of TestPrep®. There is some overlap in perceptions that that administration had a clear plan for data use, modeled data use and provided resources for data use which would indicate that they are not statistically different. However, there is no overlap between administration encouraged data use and administration modeled data use and provided resources for program use. Thus, there is likely a statistically significant difference between encouraging the use of USA TestPrep® and providing resources and support for program usage and modeling data usage.

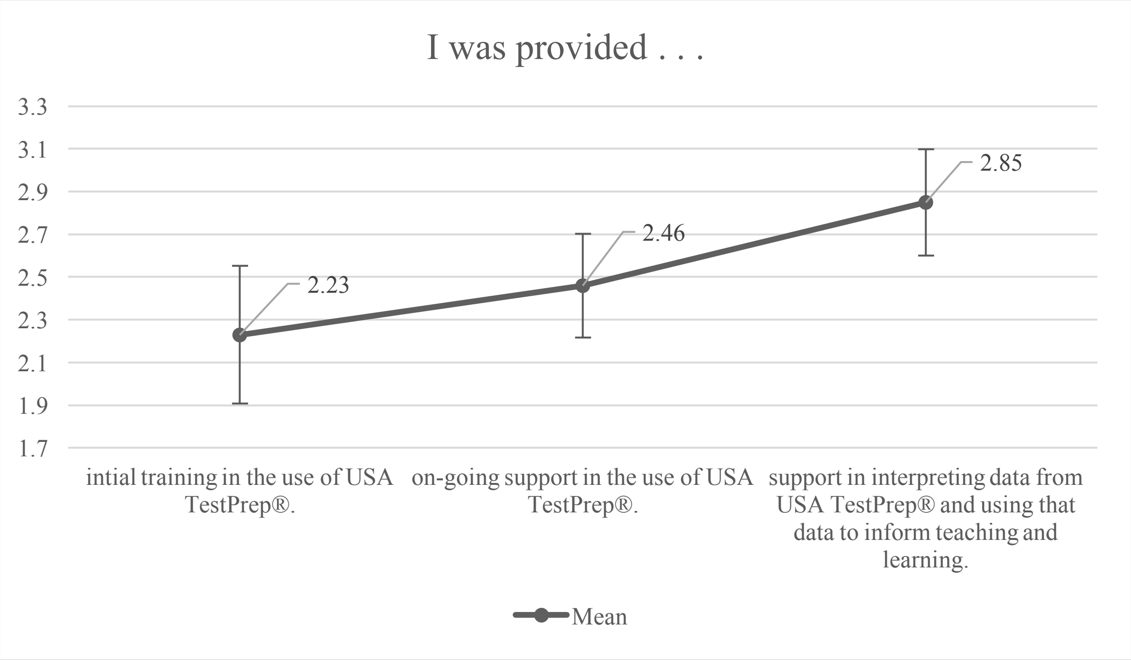


***Figure 3*. Descriptive Statistics for Providing Resources for Use of USA TestPrep$^{®}$. The bars above and below the mean depict a 95% confidence interval for the estimate of the mean.**

In Figure 3, there is some overlap in the confidence intervals for receiving both initial training in program use and on-going support for program use. There is less overlap between on-going support for program use and the confidence intervals for providing support in using data to inform teaching and learning. However, it should be noted that all three still indicate that teachers do not perceive that they were provided support for any of the three and the overlap among the three indicates that there is no statistical difference among them.

In Figures 4 and 5, the researcher illustrates prompts that were designed to measure perceived program utility. Eighty-five percent of participants reported that USA TestPrep$^{®}$ was user-friendly and that it also provided useful data.  Ninety-two percent of respondents indicated that the program provided tests and activities with higher order thinking skills. Yet, only 54% of respondents indicated that the program helped them guide learning and instruction.

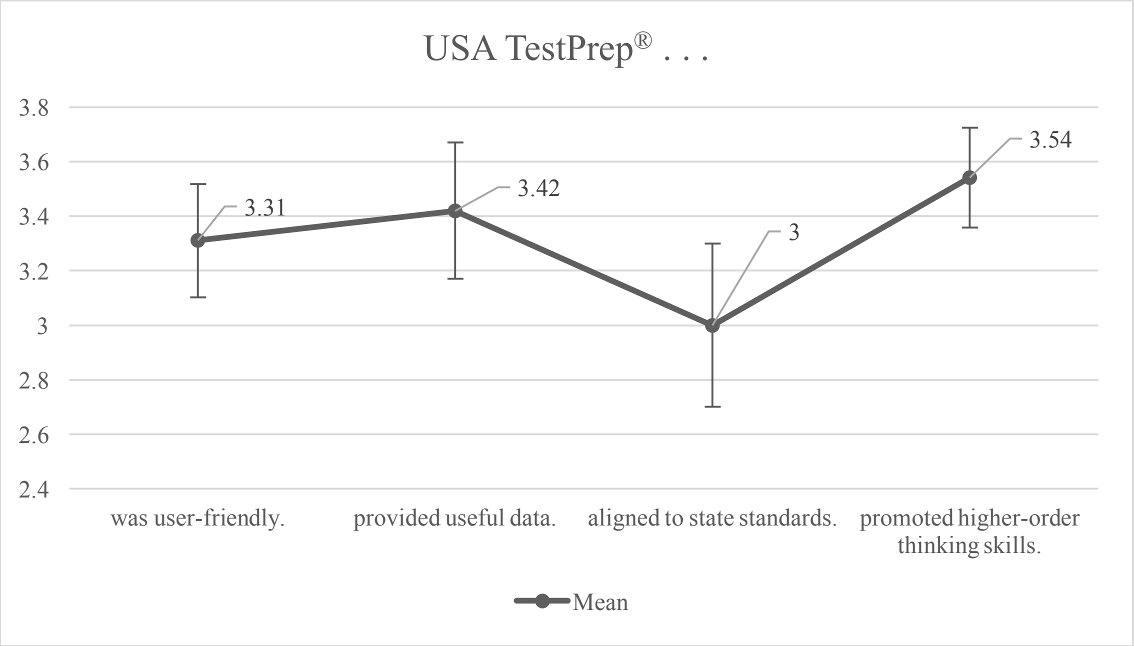


***Figure 4*****. Descriptive Statistics for Utility and Alignment of USA TestPrep®. The bars above and below the mean depict a 95% confidence interval for the estimate of the mean.**

Eighty-five percent of participants indicated that USA TestPrep® measured academic knowledge. Teachers were less confident in program ability to measure deeper learning.  Seventy-seven percent agreed that the program measured critical thinking skills as well as its ability to measure problem-solving skills.  Only forty-six percent of responses indicated agreement that USA TestPrep® measured creativity.

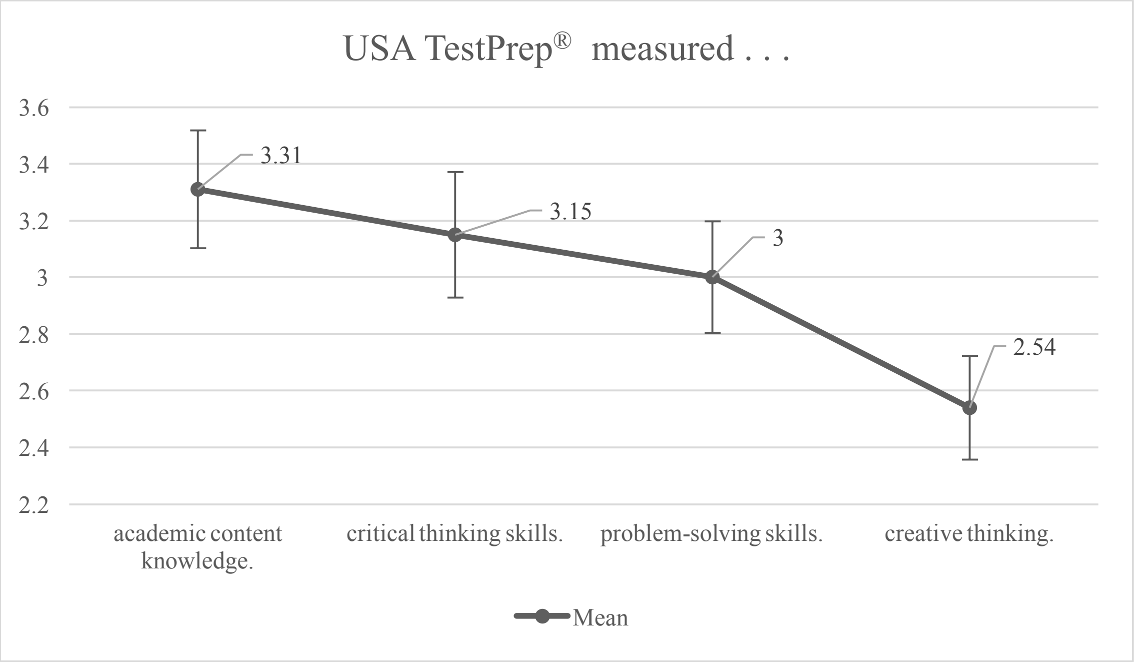


***Figure 5.* Descriptive Statistics for USA TestPrep® Measures of Deeper Learning. The bars above and below the mean depict a 95% confidence interval for the estimate of the mean.**

Confidence intervals depicted in Figures 4 and 5 reflect particularly wide confidence intervals. This is due, in part, to the small sample size. The only confidence intervals in Figure 4 that do not overlap are those for USA TestPrep® is aligned to state standards and USA TestPrep® promoted higher-order thinking skills indicating that there is likely a statistical difference. In Figure 5 there is an outlier. Descriptive statistics for teachers' perception that USA TestPrep® indicate that in regard to creative thinking both the mean and confidence levels fall below and outside of those for the program's ability to measure academic content knowledge and other components of deeper learning such as critical thinking skills and problem-solving skills. The lack of overlap with these three and creative thinking skills suggests that there is a likely statistical significance between perceptions of USA TestPrep® measurement of creative thinking and its measurement of critical thinking, problem solving, and academic content knowledge.

The survey also included two free-response portions. All thirteen participants answered the first of the two free response questions. In answering one survey item, teachers were asked how they created their benchmark assessments. Did they use program-generated benchmark tests that mimicked OCCT blueprints or if they used customized benchmarks that were created by teachers selecting question items based on standards and objectives covered in class and why? All thirteen responses reported using customized benchmark assessments created by individual teachers or departments.

Two teachers believed that the program-generated benchmarks did not meet their standards. Thus, they customized the tests. One teacher mentioned that customizing the benchmark assessment made it “more appropriate for my individual student needs.” One eighth grade teacher referenced “breadth of knowledge” when describing how the math benchmarks were created. The average question difficulty had to come out to be a “breadth of knowledge” level two or higher to be given as a benchmark assessment. Two teachers described selecting their own question items but doing so in a manner that mimics the OCCT or state test blueprints. One of those two described how they adjusted the benchmark tests over a three-year period until the benchmark test became, “extremely predictive of a student’s performance on the state test.”

In answering a second free response item, teachers were asked to describe how they used student benchmark testing results.  Twelve of the survey participants responded.  Seven of the twelve responses referenced identifying either specific objectives that had lower levels of mastery or individual students with low test scores. Six of those seven also referenced using that data to remediate or re-teach. Two of the seven referenced using that data to not only identify weaknesses but strengths as well.

Two respondents described using the data for further instruction, but did not reference weak areas of mastery in regard to objectives or students nor did they mention remediation. Two of the twelve responses referenced evaluating progress throughout the year with one of the two describing using the compared scores to set goals.  A final response referenced using benchmark results to "encourage my students and competition between classes."

### *Analysis of Focus Group Data*

Three teachers participated in the focus group. The researcher designed the focus group protocol to generate a discussion regarding teachers' beliefs and perceptions of USA TestPrep® as a tool to guide teaching and learning. The protocol is included in Appendix B.

The researcher identified a key pattern that emerged from the discussion among participants. Although the protocol questions never mentioned or made reference to benchmark testing, all participants primarily referenced USATestPrep® in terms of testing.  When asked how they would describe USATestPrep®, even though benchmark assessments are just one component of the program all three described the program in terms of benchmark testing. When asked what the first thing that comes to mind upon hearing USATestPrep®, once again all three participants referenced benchmark testing. One participant responded, "I think of just the standardized testing and sometimes that strikes fear in a student." Throughout the course of the focus group discussion, excluding the name of the program, USATestPrep®, participants used the word test or testing twenty times and the term benchmark twelve times.

The process of creating benchmark exams was the subtheme most often used when referencing benchmark exams. Teachers discussed the process of selecting questions and creating the test. English and math teachers described picking questions for benchmarks based on objectives covered in class that six weeks what they thought students would understand and had seen before. Both an English and math teacher also reported selecting questions based upon depth of knowledge and using that component to create a test that reflected the depth of knowledge on the state assessment. Although the math teacher added that there were sometimes not enough questions at the desired depth of knowledge level by the time the final benchmark approached. However, all teachers in the focus group spoke favorably of the choice they had in the process of creating benchmark assessments.

Another common subtheme was the use of data provided by the benchmark assessments and how teachers used that data. All participants found the data provided by the benchmark assessments to be a helpful tool in instruction. Other minor subthemes referenced, but not repeatedly, by teachers included procedural struggles with benchmarking and comparison to the new benchmark assessment program recently adopted by the district.

By comparison, teachers referenced other USA TestPrep® program components far less frequently. Of those components, teachers referenced games three times and other program "resources" twice. All participants spoke favorable of these components and referenced using them on occasion. As one participant said, "I did find that there were a lot of resources on USA TestPrep® that I simply did not have the time to use . . . had I had the time, I would have used those resources."

Also, important to note is the general attitudes and feelings of participants regarding USA TestPrep®. Teachers used terms such as like, thirteen times, and enjoyed four times, as opposed to don't, twice, and frustrating, five times. Additionally, focus group members discussed in more detail and at greater length what they liked about USA TestPrep® than what they found frustrating. Focus group discussion frequently centered on the favorable opinions of participants regarding the data provided by USA TestPrep®. Teachers used phrases such as "loved the data," "appreciated the data," "enjoyed using data," and "(it) was very helpful to have data in front of me." Teachers also voiced their enjoyment of choice in creating the benchmark tests and how the program "became a pretty good tool" for instruction and helping students.

Yet, the teachers were not without frustration when it came to program usage. Teachers expressed frustration that at times they had difficulty duplicating tests when, for example, they wanted to give the same test for the last benchmark of the year as the first benchmark of the year. Additionally, teachers mentioned that found that by the time they were creating the final benchmark of the year there seemed to be some shortage of options. A reading teacher mentioned a "limited" number of stories to select from and a math teacher mentioned that sometimes there were not enough questions at the desired depth of knowledge level.

Focus group members all acknowledged frustration in reference to the implementation process and not the program itself. One teacher mentioned, and others nodded in agreement, the frustration of having to "schlep our students up to the computer lab and back down again." Later in the discussion, a second teacher again referenced the frustration at having to move students to the computer lab.

Yet, while discussing their frustrations, two of the three teachers also made a positive reflection regarding the program. One teacher clarified that so it was not really the program itself that caused the frustration, but the way the process was organized. Concluding, "I can't think of anything that frustrated me." Another teacher while discussing frustration with the number of story options also acknowledged appreciation for teacher choice in selecting stories.

The focus group was designed to elicit teachers' use and perceptions of USA TestPrep® as a tool to guide DDDM. A concluding comment from the discussion best encapsulates teacher thoughts and perceptions not only because of the respondent's comments, but also in that the other participants became animated as they nodded their heads in agreement and even interrupted to echo the respondent's thoughts. When asked if there was anything else he/she would like to add, the participant began, "There is a saying that says you don't know what you've got till it's gone." The other participants laughed and nodded in agreement as the respondent continued to describe the new program that has recently replaced USA TestPrep® in the district:

> (The new program) doesn't give us the freedom of setting up to see what our kids have been through, doesn't give us the freedom to see what questions were actually missed by our kids. We will get the grades; we will see which objectives they scored best and worst at. But we just won't get the same feel. It's almost like receiving standardized test results. …but now that we have to do

> the same thing and we don't get to see that data and those results it is even more frustrating.

While the participant was comparing the two programs, the comments are a good summary of the focus groups thoughts and perceptions of USA TestPrep®. Teachers see the program primarily as a testing tool. While the process of benchmarking could be frustrating, they appreciate the freedom and choice that teachers had in creating the benchmark tests, they liked the data they had access to and how they could use it to better help students.

### *Summary*

A comparison of the two data sources provides an informative summary and more complete picture of research question three findings. The survey responses and focus group discussion yield both similarities and differences in their results. In regard to the various program components, there were more similarities in what program components focus groups and survey respondents did not use. Focus group discussion revealed that the other program components were seldom if ever used and lack of time was cited as the reason why. "I did find that there were a lot of resources on USA TestPrep® that I simply did not have the time to use. …had I had the time, I would have used those resources."

Quantitative and qualitative data from the survey echoes the focus group discussion. As found in Table 9, the majority of participants indicated that they never used the majority of program components. In fact, using benchmark assessments and student data reports were the only two components that a majority of teachers reported using. Qualitative data from the survey further supports findings from the focus group.

In a free response item, the most frequent reason given by participants for not using program components was lack of time.

Focus group discussion revealed that participants used USA TestPrep® primarily as a benchmark assessment tool. Quantitative data from survey response items supports this as well. The frequency analysis found in Table 9 shows the response item with greatest consensus in regard to what component participants used occurs when eleven of the thirteen respondents indicate using benchmark tests one to three times per semester.

Closely related, focus group discussion revealed that teachers not only used the data provided from benchmark assessments but valued the data as well. This is evidenced by statements such as “loved the data,” “appreciated the data,” “enjoyed using the data,” and “It was helpful to have the data in front of me.” Once again, quantitative data from the survey supports the findings that teachers used data produced from USA TestPrep®. According to survey responses in Table 9, the second most commonly used component of USA TestPrep® was student data reports.

Yet, here is where the similarities with data usage end. As mentioned, focus group discussion revealed that participants not only used the data but *valued* the data. Participants repeatedly referenced not only using the data but specifying using it to review and remediate and at times provide enrichment at home. While according to survey findings, the student data reports component was the second most used component, only eight of the thirteen survey responses reported using the student data reports.

When considering other survey response items related to data use the researcher finds that the results are mixed. Ninety-two percent of survey responses agreed that

USA TestPrep® was adopted to improve teaching and learning while eighty-five percent agreed that the program provided useful data. However, the disconnect begins to appear with other data-usage related response items.  Seventy-seven percent of survey participants agree that USA TestPrep® was adopted to provide enrichment opportunities. Only fifty-four percent of respondents agreed that USA TestPrep® helped guide instruction and learning. A mere thirty-nine percent of respondents indicated that they believed the program helped prepare students to be college and career ready.

In regard to helping guide the teaching and learning of specific populations, the disconnect continues. Survey respondents overwhelmingly disagree with the ability of USA TestPrep® to do so. Only forty-six percent of respondents agree that USA TestPrep® was adopted to help serve high-risk students and only twenty-three percent agree that USA TestPrep® was adopted to serve students with cognitive disabilities. The disconnect with using USA TestPrep® to guide instruction and learning will be addressed further in the discussion.

.

# Chapter 6: Discussion

## Overview of Study

College and career readiness (CCR) is a pivotal outcome of recent legislation such as *Race to the Top* initiative and the Common Core State Standards (CCSS) movement (Conley, Drummond, de Gonzalez, Rooseboom, & Stout, 2011; Lombardi, Conley, Seburn, & Downs, 2012). Despite the demands of recent federal and state policy, many students are not proficient in the skills necessary to meet the demands of college and careers (Conley & Darling-Hammond, 2013; Bitter & Loney, 2015).

USA TestPrep® is an online software program designed to offer test preparation resources to students, teachers, and schools. The purpose of this study was to consider the efficacy of USA TestPrep® in broadening teacher capacity to meet the cognitive learning needs of students and develop student skills and knowledge so that they are college and career ready (CCR). To fulfill this purpose, the researcher utilized a mixed-methods approach to this study via the following research questions:

1. Do USA TestPrep® benchmark assessments predict Oklahoma Core Curriculum Test performance?
2. Using Webb's Depth of Knowledge Framework, how well do the USA TestPrep® benchmark assessments align with the cognitive dimension of deeper learning?
3. How did teachers use and perceive the program to assist them in improving instruction and student learning?

In order to understand program efficacy in helping teachers prepare students to be college and career ready, the researcher first examined the nature of the relationship between USA TestPrep® benchmark assessment results and Oklahoma Core Curriculum Test (OCCT) results. Quantitative data from two continuous variables, USA TestPrep® benchmark assessment results and OCCT scores were used to run a Pearson's correlation coefficient analysis.

Once the nature of the relationship between USA TestPrep® benchmark assessment results and OCCT results was examined, the study then determined whether or not those USA TestPrep® benchmark tests assessed student deeper learning as measured through Norman Webb's depth of knowledge. Qualitative data from the USA TestPrep® benchmark assessments was used to conduct a content analysis to determine the number of assessment items from each benchmark assessment given during the course of the study that fell at each level of Norman Webb's depth of knowledge.

Finally, the researcher sent recruitment letters via email to teachers asking them to participate in a survey and/or focus group regarding their use and perceptions of USA TestPrep® as a tool to aid in developing students who are college and career ready. Quantitative data included descriptive statistics drawn from Likert-type questions. An analysis of themes was conducted from qualitative data collected from the free-response portion of the survey as well as responses to the focus group discussion.

## Discussion of Key Findings

### *Research question one*

The researcher began the study with research question one: an examination of the nature of the relationship between seventh and eighth grade math and reading USA TestPrep® benchmark assessments and OCCT results. The researcher found that there is a positive correlation between student scores on all forty-five benchmark tests and their corresponding OCCT. All results were statistically significant and indicative of a strong, highly significant positive relationship.

Despite the fact that the Pearson's *r* coefficient values for all five subjects were indicative of a strong, highly significant positive relationship; some of the *r* values were quite high, others were not, and none were an absolute one. Benchmark scores cannot perfectly predict OCCT results. Thus, the results are tempered when considering the $r^2$ values. An $r^2$ value explains how much of the variance in the dependent variable can be explained (Vogt, 2007). At the high end of the results, fifty percent or less of the variance in OCCT results is predicted by benchmark performance. What phenomena account for the rest of the variance in OCCT performance?

Two issues relating to benchmark creation might explain the variance. The Pearson's correlation coefficients did not demonstrate increasing *r* value from one benchmark to the next in any given year for any of the five courses included in this study. One could expect that as the instructional year progressed, student mastery of skills would increase and thus would reflect in increasing *r* values from one benchmark to the next. In fact, this is not the case.

This trend of seemingly indiscriminate Pearson's *r* coefficient values suggests that perhaps there is an issue with the breadth of content on the benchmark assessments. This is perhaps due to a misalignment of categorical concurrence between the standards and benchmark assessments. Categorical concurrence is one of four categories that measure the alignment assessments to standards. Categorical concurrence occurs when similar standards and objectives content are present in assessments and standards (Webb, 1999). A discussion of this will be fleshed out further in the discussion on research question three regarding teachers' use and perceptions of USA TestPrep® as a tool to assist them in improving instruction and student learning.

Secondly, remaining variance might also be explained by not only the categorical concurrence but by the depth of knowledge of question items as well. As discussed in the conceptual framework chapter, depth of knowledge appertains to cognitive requirements and complexity of tasks (Herman et al., 2007; Webb, 1997) and was created to examine the relationship between assessments and expectations (Webb, 1999; 2006; 2010). As reported in the findings for research question two, only two percent of all benchmark assessment items collected as part of the study were at level three depth of knowledge.  The remaining ninety-eight percent were level one and two items. According to Test and Item Specifications, the distribution range for level three on both seventh and eighth grade reading and math OCCT exams is fifteen to twenty-five percent. Student performance on depth of knowledge level three questions on the OCCT exams could explain some of the remaining variance. As created, the benchmark assessments provide insufficient evidence to measure student performance on depth of knowledge level three question items. As such, the benchmarks cannot accurately measure or predict student performance on these types of question items. This is further addressed in the discussion on research question two regarding the alignment of the cognitive dimension of deeper learning with benchmark assessments.

While the benchmark assessments can be a useful aid in predicting OCCT performance, additional variance could be explained by factors other than those relating to benchmark creation.  The researcher suggests that other factors such as teacher perception and use of the program could account for the remaining variance. As reported in the findings for research question three, teachers' perceptions of the benchmark process and having to "schlep students up to the computer lab" was not favorable. Furthermore, their perceived support for program resources and training were low, with the mean and confidence intervals both falling within the range of disagreement. Not within the scope of the study, variables such as student perceptions of benchmark testing and OCCT testing might also account for the remaining variance. Did students respect the benchmark exam and put the appropriate effort into completing the assessments? How did their focus and effort and perceptions of the importance of the benchmark exam compare to the OCCT? While this study does not consider student perceptions of either benchmark testing or OCCT exams and how their perceptions influence performance on either one of the assessments, those perceptions might account for some of the variance as well.

### *Research question two*

The researcher measured how well the benchmark assessments aligned with the cognitive dimension of deeper learning as seen through the conceptual framework of Norman Webb's depth of knowledge. According to Herman and Linn (2013), depth of knowledge levels 3 and 4 are important components of deeper learning because they require students to demonstrate critical thinking skills such as applying and synthesizing. In this study, the benchmark assessments' alignment with levels three and four of Norman Webb's depth of knowledge is crucial in determining the program's role in helping teachers facilitate deeper learning in their students.

By far, both math and reading benchmark assessments did not align with the Test and Item Specifications distribution ranges for depth of knowledge levels. While their distributions may fall below or exceed the distribution range, reading benchmark assessments most closely mirrored the OCCT blueprints. Only two of the eighteen reading benchmarks did not have a single question for a depth of knowledge level three. By contrast, math had fourteen benchmark assessments that failed to address a depth of knowledge level three. Neither subject assessed at level four depth of knowledge due in large part to the fact that USA TestPrep® program design did not include level four questions as part of the benchmark assessment question banks.

Other researchers have found similar results regarding assessing depth of knowledge. A study of four states' mathematics assessments found that a high percentage of their assessment items fell below the depth of knowledge level for their corresponding objective (Webb, 1999). Lindberg, Shibley Hyde, Petersen and Linn (2010) in their meta-analysis of state mathematics assessments found that level three and level four depth of knowledge items were markedly lacking on mathematics assessments. They argue that this failure to emphasize critical thinking and complex problem solving not only leaves an incomplete and inaccurate measurement of student mathematics skills, it also fails to emphasize the skills that our society needs. The assessment systems of many states are unable to identify and describe students who are college and career ready (Darling-Hammond & Adamson, 2013; Darling-Hammond & Conley, 2015).

With policies that place high importance on developing CCR, the ability to measure deeper learning as defined by deep content knowledge, critical thinking, and complex problem-solving (NRC, 2011) resulting in student proficiencies such as the ability to think creatively, formulate and solve problems, and data analysis and statistical reasoning (William and Flora Hewlett Foundation, 2013) must be present in student assessments. The researcher found in the analysis of research question two, the benchmark assessments did not provide a sufficient measurement of deeper learning. The vast majority of assessments did not measure student strategic or extended thinking.

The researcher suggests that this is due to benchmark creation issues. In part, this is due to program limitations in the amount of level three and four questions as evidence in focus group responses. However, all of the survey responses indicated that teachers did not use program generated benchmarks designed to mimic OCCT assessments in both breadth and depth of content and depth of knowledge. Rather, most benchmarks were created by teachers who individually selected test items based upon standards covered thus far during the academic year and by what they deemed appropriate for their students. Considering this lack of program-generated benchmark tests and the small sample size of thirteen survey respondents, it cannot be definitively concluded whether the insufficient measure of deeper learning is due to program design or program implementation.

In and of themselves, assessments of deeper learning can encourage desired changes in instructional practice (Faxon-Mills, Hamilton, Rudnick, & Stecher, 2013). In this case of this study where there is a decided lack of assessments that address deeper learning, the benchmarks as created using the USA TestPrep® are not a tool which guides DDDM which facilitates preparing students who are college and career ready. In the discussion on research question three, the researcher will further address the alignment between the benchmark assessments and the cognitive dimension of deeper learning.

### *Research question three*

In studying research question three, the researcher attempted to gain insight into teacher use and perception of USA TestPrep® as a tool to guide DDDM. Both survey results and focus group discussion revealed that teachers primarily use the program as an assessment tool used for predictive purposes and to a lesser extent as a tool to guide instruction.

In the discussion on research question one, the researcher suggested that the seemingly indiscriminate relationship between benchmark assessments and OCCT performance was due to a lack of categorical concurrence.  An analysis of teacher survey responses and focus group discussion supports this conclusion. In the free-response portions of the survey and the focus group discussion, respondents indicated that they customized the benchmark tests based on current standards.  Some specified that they selected questions based upon topics that had been covered or ones that were more appropriate for their individual student needs. Based upon study findings, benchmark exams were reflective of the content studied during the six-week period leading up to the benchmark assessment and not the entire breadth and depth of the standards and objectives that would be covered on the OCCT.

In the discussion of research question two findings, the researcher asserted that the benchmark assessments did not adequately align with the cognitive dimension of deeper learning. The researcher finds that focus group comments support this when teachers mention considering depth of knowledge when choosing questions but finding the program lacking in adequate questions at the different levels of depth of knowledge.

But, that is not to say that there was necessarily a lack of consideration for depth of knowledge. Teacher survey responses suggest that perhaps there is a misalignment in their perceptions of depth of knowledge use and their actual use of depth of knowledge. While the frequency analysis results for the prompt teachers perceived USA TestPrep® to measure higher-order thinking skills, the mean score dropped and the standard deviation increased when teachers were asked the perceptions regarding USA ability to measure academic knowledge, critical thinking skills, problem-solving skills, and creativity. This suggests that there is either a differing definition of these terms as evidenced in a teacher's use of the term "breadth of knowledge level 2" in the free response portion of the survey or differences in identifying questions that measure these dimensions of deeper learning or perhaps content standards were more a more central focus in item selection than was depth of knowledge. As stated earlier, the majority of benchmark tests were created by individual teachers who selected individual assessment items based upon content standards covered what they thought was appropriate for their individual student needs.

The crux of research question three is how USA TestPrep® was used as a tool in data driven decision-making. Perie et al. (2007) assert that benchmark assessments can serve three purposes: predictive, instructional, or evaluative. Teachers perception that USA TestPrep® was adopted to predict OCCT performance and measure academic content knowledge had higher mean scores than did purposes which would have instructional or evaluative intent.

The next highest perceived purpose was instructional purposes: for improving teaching and learning and closing the achievement gap. Both of which had a mean score of three, nearly a half of a point lower on a four-point scale – a significant difference. Furthermore, the confidence intervals on the upper end extended into somewhat agree category and the lower end into the somewhat disagree category. This is indicative of less consensus among respondents. The lack of consensus for instructional purposes is further evidenced by the reported perceptions that USA TestPrep® was not adopted to better serve students with disabilities or better serve high-risk students. Their means scores were amongst the lowest as was preparing CCR. The mean scores for evaluative purposes such as evaluating programs and interventions and evaluating teachers were low, representative of not perceiving program purposes to be evaluative in nature.

Thus, the researcher suggests that teachers primarily perceive the intent of the benchmark assessment system as fulfilling predictive purposes. The program has many components other than the benchmark testing feature.  Yet, survey responses and focus group discussion reveal these components are rarely used and USA TestPrep® is nearly always framed in a testing context. While teachers favorably view the data provided by the program and do reference using the data, these comments fell secondary to discussions focusing on creating the test and the testing process. Teachers do not mention how they *change* instruction or use data to create opportunities for deeper learning.

Other researchers have found similar results.  Blanc et al. (2010) found that Philadelphia teachers responded to the need for improved test results by using benchmarks to predict student outcomes on standardized testing. As stated, teachers did report using the data to primarily to identify student and objective weakness, but also strengths. Oláh, Lawrence, and Riggan (2010) in an article based on the same study as Blanc et al. found that Philadelphia teachers used benchmarks to identify areas for emphasis. They argue that more significant than do teachers use data is *how* teachers use benchmark data. Shepard, Davidson, and Bowman (2011) found similar results with their study of middle school math teachers. They concluded that interim assessments primarily were used to measure progress and predict performance on standardized assessments and while teachers expressed a desire to use data in instruction there was little professional development to help them do so.

In regard to this study, mean scores from survey response items related to program support are indicative of the fact that teachers perceived that they did not receive support to either use the program and/or use the data generated from the program. This conclusion is corroborated by focus group responses that indicated teachers valued and appreciated the data, used it to identify strengths and weaknesses, but did not relate how they used the data to change instruction. Both of which would perhaps help indicate why teachers did not perceive the program to have instructional purposes that ultimately lead to extending teacher capacity to meet the learning needs of students and develop student skills and knowledge so that they are college and career ready.

In summary, the intent of this study was to examine the efficacy of USA TestPrep® in extending teacher capacity to meet the learning needs of students and develop skills and knowledge so that students are ready for college and career. The first step in preparing CCR is to understand if the program measures what it purports to measure. While the strong, positive correlations indicate that benchmark assessments created using the USA TestPrep® program were predictive of OCCT performance, it cannot be said that the program as implemented in the district included in the study measures what it claims to measure. The benchmark performance accounts for too little of the variance in OCCT performance.

Yet, the intent of the study was to go beyond predictive purposes of the program to examine how it influences instruction and learning. While teachers did make mention of using data from the program to guide instruction, they did not articulate how they did so. Furthermore, they did not perceive that the program helped them develop students who were college and career ready. Teachers' confidence in the capacity of the program to provide benchmark tests with higher order thinking skills was variegated and, as a whole, teachers were much less confident when they were asked about specific components of deeper learning. Additionally, results suggest that teachers did not see the opportunities to foster other deeper learning domains such as the intrapersonal. USA TestPrep® program design is such that students may move through self-guided activities that allow the individual student to work through components such as video lessons, games, practice questions and even practice benchmark tests. As indicated by survey results regarding the usage of program components and the focus group discussion, teachers did not use the program to foster intrapersonal domain student proficiencies such as initiative and self-regulation.

This indicates that perhaps teachers' orientations to deeper learning are not strong. The misalignment and variegated responses to not only components of the cognitive domain of deeper learning, but the lack of addressing the intrapersonal domain suggests that they do not have a conceptual understanding of deeper learning or its components. Emphasis continued to return to testing features and less on how the program influenced teaching and learning. Perhaps the name of the program itself, USA TestPrep®, shaped perceptions of its purpose. It is a marketable "test prep" tool designed for states and local districts that look for ways respond to the demands of policy and high-stakes testing. Yet, in the case of this study, while there is data, it does not drive instructional decision-making to help produce students who are college and career ready.

## Limitations and Implications for Future Studies

One potential criticism of the study is that it is limited in its scope and sample of middle school teachers. The scope of the study is limited to one suburban middle school. The well-designed nature of the study with its mixed methods approach that allows for triangulation of data helps ensure that the findings are nonetheless valid. Broadening the sample size to all middle schools in Oklahoma that use the program is left to future studies. Future studies should include not only other suburban schools but urban and rural schools and thus extend the findings to be more representative of the population of schools, teachers, and students using USA TestPrep® and taking OCCT assessments. Furthermore, this study was a bivariate correlational analysis. A multivariate analysis would allow the researcher to consider how other variables may moderate or mediate the program's ability to help teachers provide deeper learning opportunities and develop CCR.

Additionally, the program is limited in its focus on the cognitive dimension of deeper learning. The first step in program evaluation should studying rather or not it measure what it intends to measure: the cognitive dimension.  But deeper learning encompasses three dimensions: the cognitive dimension as well as the interpersonal and intrapersonal dimensions. Future studies should consider to what extent the program facilitates opportunities for deeper learning in the interpersonal and intrapersonal dimensions, particularly by including an analysis of the components that allow for student-guided instruction and assessment. According to the Gordon Commission, the best assessments facilitate the mastery of concepts if they allow students to assess their own progress (as cited by Darling-Hammond et al., 2013, p. 2).  Extending the study to these features of the program usage will further the findings of this study and allow for a more complete analysis of how USA TestPrep® provides opportunities for deeper learning.

Not within the scope of this study is USA TestPrep® as an intervention tool. This study did not set out to consider whether or not scores increased with the use of USA TestPrep®, only the nature of the relationship between benchmark assessment scores and OCCT. As such, while the correlations were indicative of a positive relationship in each of the three years of the study, they do not indicate if OCCT performance increased or decreased from year to year of the study. Conceptually, scores could decrease each year of the study and a correlational analysis could still indicate a strong positive relationship between benchmark assessments and OCCT performance.

Considering the extent to which USA TestPrep® serves as an intervention tool for increasing student performance would allow the researcher to consider whether or not schools and students continue to demonstrate increasing performance with each year that they use USA TestPrep®. Such an approach would extend the findings of this study to consider the capability of other courses in the USA TestPrep® program. Additionally, in an era of reduced financial support for schools, districts must be judicious in how they distribute their fiscal resources. Information regarding the capacity of the program as an intervention tool would valuable to schools using and considering using the program.

The focus group included three participants. While scholars such as Krueger and Casey (2000) and Peak and Fothergill (2009) support the utility of smaller sized groups, the size of the focus group is a limitation of this study. Including a greater number of teacher participants in the focus group would have provided more data to confirm and triangulate study results and/or identify areas for future research. Future studies should seek to recruit more participants in focus group discussions.

Finally, this study is limited in that it only considers teachers' uses and perceptions of USA TestPrep®. Adding a district-level component to future studies could also provide a broader understanding of program effectiveness in helping teachers provide opportunities for deeper learning. There is a reciprocal relationship between district/school policies and practices and teacher beliefs and knowledge (Faxon-Mills et. al, 2013, p.7). Future studies should consider how district level policies and practices might not only mediate teacher perceptions but providing deeper learning opportunities to students and preparing them for college and career.

# References

About Us. (2018). Retrieved from https://www.usatestprep.com/about-us

Adams, C. M., Forsyth, P. B., Ware, J, Mwavita, M. (2016). The informational significance of A-F school accountability grades. *Teachers College Record, 118*(7), 1-31.

Adams, C. M., Ford, T. G., Forsyth, P. B., Ware, J. K., Olsen, J. J., Lepine, J. A., Sr., et al. (2017). Next generation accountability: A vision for improvement under ESSA. Palo Alto, CA: Learning Policy Institute.

Amrein, A. L., & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), 1-74. Retrieved from: http://epaa.asu.edu/ojs/article/view/297/423

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287-302. http://dx.doi.org/10.1080/0305764X.2011.607151

Bitter, C., & Loney, E. (2015). Deeper learning: Improving student outcomes for college, career, and civic life. Washington, DC: American Institutes for Research. Retrieved from: http://educationpolicy.air.org/sites/default/files/Brief-DeeperLearning.pdf

Bitter, C., Taylor, J., Zeiser, K. L., & Rickles, J. (2014). *Providing opportunities for deeper learning. Report 2 findings from the study of deeper learning: Opportunities and outcomes*. Washington, DC: American Institute for Research.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi delta Kappan 86*(1), 9-21.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, *5*(1), 7-74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K.E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education 85*(2), 205-225. http://dx.doi.org/ 10.1080/01619561003685379

Campbell, D.T. (1976). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67-90.

Comprehensive Solution. (2018). Retrieved from https://www.usatestprep.com/comprehensive-solution

Conley, D. T., Drummond, K. V., de Gonzalez, A., Rooseboom, J., & Stout, O. (2011). *Reaching the goal: The applicability and importance of the Common Core State Standards to college and career readiness*. Eugene, OR: Educational Policy and Improvement Center. Retrieved from: http://files.eric.ed.gov/fulltext/ED537872.pdf

Conley, D. T., & Darling-Hammond, L. (2013). Creating systems of assessment for deeper learning. Stanford, CA: Stanford Center for Opportunity in Policy in Education. Retrieved from: http://scee.groupsite.com/uploads/files/x/000/09e/76f/creating-systems-assessment-deeper-learning.pdf.

Creswell, J. W. & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage.

Darling-Hammond, L. (2007). Third Annual Brown Lecture in Education Research: The flat earth and education: How America's commitment to equity will determine our future. *Educational Researcher, 36*(6), 318-334. Retrieved from http://www.jstor.org/stable/30133808

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 Center for Opportunity Policy in Education. Retrieved from: https://edpolicy.stanford.edu/sites/default/files/publications/developing-assessments-deeper-learning-costs-and-benefits-using-tests-help-students-learn_1.pdf

Darling-Hammond, L. & Conley, D. T. (2015) Assessments systems for deeper learning. In J.A. Bellanca (Ed.), *Deeper learning: Beyond 21st century skills* (pp. 235-271). Bloomington, IN: Solution Tree Press.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aaber, L., Baker, E., . . . Steele, C. M. (2013). Criteria for high-quality assessment. Stanford, CA: Stanford Center for Opportunity Policy in Education. Retrieved from: https://edpolicy.stanford.edu/sites/default/files/publications/criteria-higher-quality-assessment_2.pdf

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives, 22(*86). http://dx.doi.org/10/14507/epaa.v22n86.2014.

Deci, E. L, Spiegel, N. H., Ryan, R. M., Koestner, R., & Kauffman, M. (1982). The effects of performance standards on teaching styles: Behavior of controlling teachers. *Journal of Educational Psychology, 74*(6), 852-859.

DuFour, R. & DuFour, R. (2015). Deeper learning for students requires deeper learning for educators. In J.A. Bellanca (Ed.), *Deeper learning: Beyond 21st century skills* (pp. 21-52). Bloomington, IN: Solution Tree Press.

Every Student Succeeds Act of 2015, Pub. L. No. 114-95, 20 U.S.C. §114 (20015).

Faxon-Mills, S., Hamilton, L. S., Rudnick, M., & Stecher, B. M. (2013). New assessments, better instruction? Santa Monica, CA: RAND Corporation.

Firestone, W. A. (1987). Meaning in method: The rhetoric of quantitative and qualitative research. *Educational Researcher, 16*(7), 16 – 21.

Ford, T. G., Van Sickle, M. E., & Fazio-Brunson, M. (2016). The role of "informational signficance" in shaping Louisiana elementary teachers' use of high-stakes teacher evaluation data for instructional decision-making. In K.Kappler-Hewitt & A. Amrein-Beardsley (Eds.), *Student Growth Measures in Policy and Practice: Intended and Unintended Consequences of High-Stakes Teacher Evaluations*. Retrieved from: https://www.researchgate.net/publication/283089389_The_Role_of_Informational_Significance_in_Shaping_Louisiana_Elementary_Teachers%27_Use_of_High_Stakes_Teacher_Evaluation_Data_for_Instructional_Decision_Making

Fuchs, L. S., & Fuchs, D. (1986). Effects of systemic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199-208.

Fullan, M. (2015) Assessments systems for deeper learning. In J.A. Bellanca (Ed.), *Deeper learning: Beyond 21st century skills* (pp. 274-284). Bloomington, IN: Solution Tree Press.

Goertz, M. E., Oláh, L. N., & Riggan, M. (2009). *Can interim assessments be used for instructional change*. Philadelphia, PA: Consortium for Policy Research in Education.

Gulikers, J. T. M., Biemans, H. J. A., Wesselink, R., & van der Wel, M. (2013). Aligning formative and summative assessments: A collaborative action research challenging teacher conceptions. *Studies in Educational Evaluation, 39*(2), 116-124.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J. A., & Wayman, J. C. (2009). Using student achievement data to support instructional decision making (No. NCEE 2009-4067). Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *The Curriculum Journal, 16*(2), 207-223.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

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

Herman, J. L. & La Torre Matrundola, D, & Wang, J. (2015). *On the road to assessing deeper learning: What direction do test blueprints provide* (CRESST Report 849). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L. & Linn, R. L. (2013). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia* (CRESST Report 823). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A case study. *Applied Measures in Education 20*(1), 101-126.

Huberman, M, Bitter, C., Anthony, J., & O'Day, J. (2014) The shape of deeper learning: Strategies, structures, and cultures in deeper learning network high schools. Findings from the study of deeper learning opportunities and outcomes: Report 1. *American Institutes for Research*.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Katona, G. (1942). Organizing and memorizing: a reply to Dr. Melton. *The American Journal of Psychology*, *55*(2), 273-275. http://dx.doi.org/10.2307/1417090

Kelley, T. L. (1927). *Interpretation of educational measurements.* Yonkers-on-Hudson, NY: World Book.

Knight, P. T. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education, 27*(3), 275-286.

Kobrin, J. L. (2007). Determining SAT benchmarks for college readiness. College Board Research Note RN-30. New York, NY: College Board. Retrieved from: http://research.collegeboard.org/sites/default/files/publications/2012/7/researchnote-2007-30-sat-benchmarks-college-readiness.pdf

Krueger, R.A., & Casey, M. (2000). *Focus groups* (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Lau, A. M. S. (2016). 'Formative good, summative bad?'–A review of the dichotomy in assessment literature. *Journal of Further and Higher Education*, *40*(4), 509-525.

Lazear, E. P. (2006). Terrorism and teaching to the test. *The Quarterly Journal of Economics, 121*(3), 1029-1061.

Lee, J. & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity and resources: State NAEP 1990-2009 reading and math achievement gap and trends. *Educational Evaluation and Policy Analysis, 34*(2), 209-231. doi: 10.3102/0162373711431604

Lindberg, S.M., Shibley Hyde, J., Petersen, J.L, and Linn, M.C. (2010). New Trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin 136*(6), 1123-1135.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29*(2)*,* 4-16

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

Linquanti, R. (2014). *Supporting formative assessment for deeper learning: A primer for policymakers*. Washington, D.C.: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/Supporting%20Formative%20Assessment%20for%20Deeper%20Learning.pdf

Lombardi, A. R., Conley, D. T., Seburn, M. A., & Downs, A. M. (2012). College and career ready assessment: Validation of the key cognitive strategies framework. *Assessment for Effective Intervention, 38*(3), 163-171.

Marsh, J.A., Pane, J.F., & Hamilton, J.S. (2006). *Making sense of data driven decision making in education: Evidence from RAND research (OP-170).* Santa Monica, CA: RAND Corporation.

Maruyama, G. (2012). Assessing college readiness: Should we be satisfied with ACT or other threshold scores? *Educational Researcher*, *41*(7), 252-261.

Mathison, S. (1988). Why triangulate? *Educational Researcher, 17*(2), 13-17.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Thousand Oaks, CA: Sage.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Committee on The Foundations of Assessment. J. Pelligrino, N. Chudowsky, and R. Glaser *Editors*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

National Research Council. (2011). *Incentives and Test-based Accountability in Education.* Committee on Incentives and Test-Based Accountability in Public Education, M. Hout and S.W. Elliot, Editors. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

National Research Council. (2012). *Education for life and work: Developing Transferable Knowledge and Skills in the 21st Century*. Committee on Defining Deeper Learning and 21st Century Skills, J.W. Pellegrino and M.L. Hilton, Editors. Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

Nichols, S.L. & Berliner, D.C. (2008). Why has high-stakes testing slipped so easily into contemporary American life? *Education Digest, 74*(4), 41-47.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 20 U.S.C. § 6319 (2002).

Noguera, P., Darling-Hammond, L. & Friedlaender, D. (2015). Equal opportunity for deeper learning. Students at the center: Deeper learning research series. Boston, MA: Jobs for the Future.

Noguera, P. (2015, October). *Equity and deeper learning*. Paper session presented at the meeting of Teaching, Learning, Coaching Conference, Denver, CO.

Oláh, L.N., Lawrence, N.R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education, 85*(2), 226-245. http://dx.doi.org/10.1080/01619561003688688

Peak, L., & Fothergill, A. (2009). Using focus groups: Lessons from studying day care centers, 9/11, and Hurricane Katrina. *Qualitative Research 9*(1), 31-59. http://dx.doi.org/10.1177/1468794108098029

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 5-13.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). The role of interim assessments in a comprehensive assessment system: A policy brief. Retrieved from: http://files.eric.ed.gov/fulltext/ED551318.pdf

Ryan, G.W., & Bernard, H.R. (2003). Techniques to identify themes. *Field Methods, 15*(1), 85-109. http://dx.doi.org/10.1177/1525822X02239569

Ryan, R. M., & Weinstein, N. (2009). Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing. *Theory and Research in Education, 7*(2), 224-233.

Schneider, A. & Ingram, H. (1990). Behavioral assumptions of policy tools. *The Journal of Politics, 52*(2), 510-529.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), Perspectives of curriculum evaluation, (pp. 39-83). Chicago, IL: Rand McNally.

Shepard, L. A. (2005) Linking formative assessment to scaffolding. *Educational Leadership, 63*(3), 66-70.

Shepard, L. A., Davidson, K. L., & Bowman, R. (2011) *How middle school mathematics teachers use interim and benchmark testing data*. (CRESST Report 897). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Stiggins, R. J. (1999). Assessment, student confidence and school success. *Phi Delta Kappan, 81*(3), 191-198.

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

Stiggins, R.J. (2015, October). *A new vision of excellence in assessment*. Paper session presented at the meeting of Teaching, Learning, Coaching Conference, Denver, CO.

Taras, M. (2005). Assessment – summative and formative – some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478.

Teacher Developed. (2018). Retrieved from https://www.usatestprep.com/teacher-developed

Vogt, W. P. (2007) *Quantitative research methods for professionals*. Boston, MA: Pearson, Education.

Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques.* Thousand Oaks, CA: Sage.

Webb, N. L. (1997). Research monograph No. 6: Criteria for alignment of expectations and assessments in mathematics and science education. Washington, D.C.: Council of Chief State School Officers.

Webb, N. L. (1999) Research monograph No. 18: Alignment of science and mathematics standards and assessments in four states. Washington, D.C.: Council of Chief State School Officers.

Webb, N. L. (2002). Depth of knowledge levels for four content areas. Retrieved August 24, 2011, from http://facstaff.wcer.wisc.edu/normw/All%20content%20areas%20%20DOK%20levels%2032802.doc

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7-25.

Webb, N. L. (2010). *Content complexity and depth of knowledge as applicable to research and practice.* Paper delivered at the ISTE International Conference on Mathematics, Science and Technology Education, Kruger National Park, South Africa, 18-21 October. Pretoria: UNISA Printers.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational evaluation, 37*, 3-14.

William and Flora Hewlett Foundation. (2013). *Deeper learning competencies.* Retrieved from http://www.hewlett.org/uploads/documents/Deeper_Learning_Defined__April_2013.pdf

Wise, L. I., and Alt, M. 2006. Assessing vertical alignment. In Aligning Assessment to Guide the Learning of All Students. State Collaborative on Assessment and Student Standards. Washington, DC: Council of Chief State School Officers.

Wyse, A.E., & Viger, S.G. (2011). How item writers understand depth of knowledge. *Educational Assessment, 16*(4), p. 185-206. http://dx.doi.org 10.1080/10627197.2011.634286

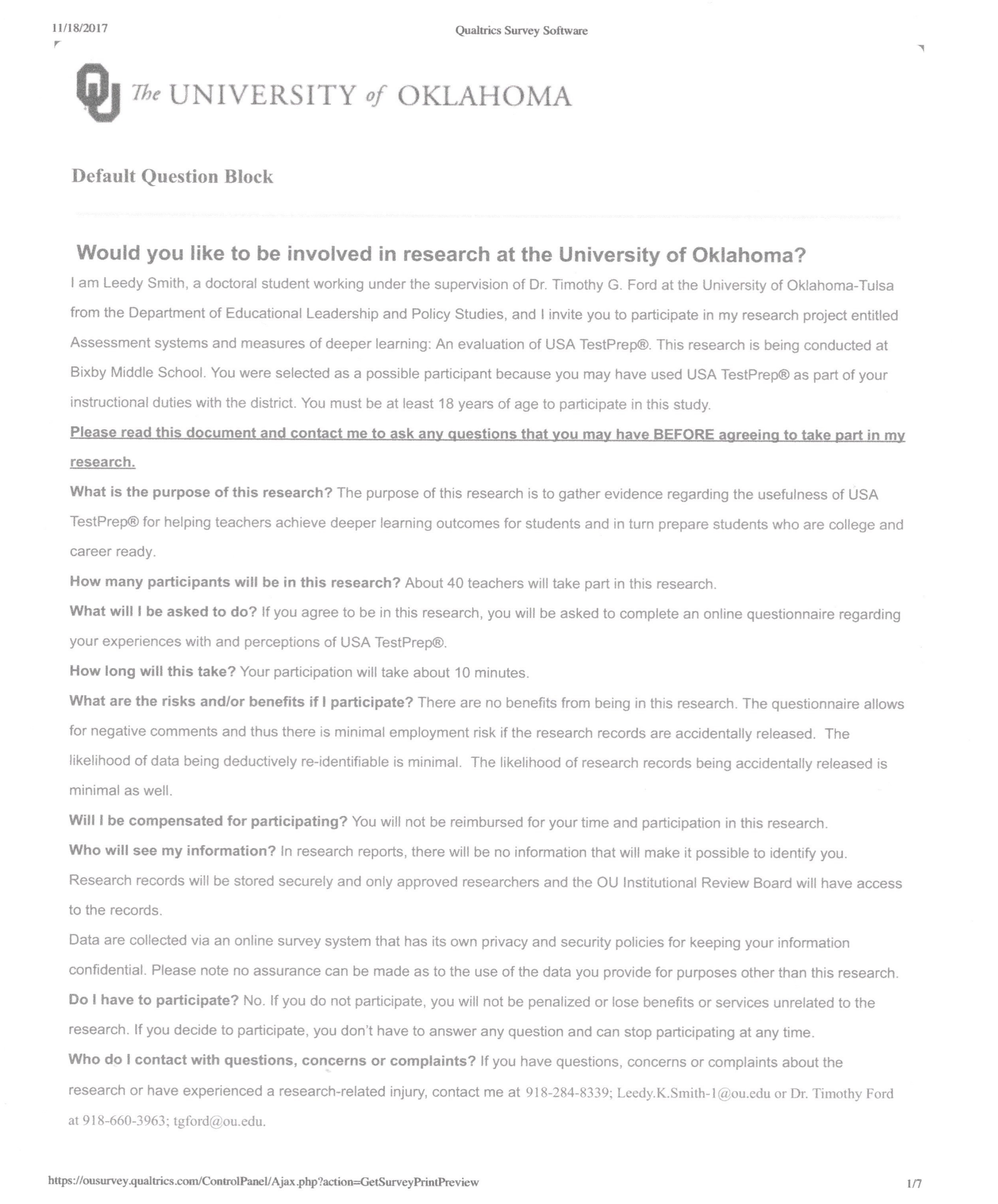
Yin, R. K. (2006). Mixed methods research: Are the methods genuinely integrated or merely parallel? *Research in the Schools, 13*(1), 41-47. Retreived from http://www.msera.org/docs/rits-v13n1-complete.pdf#page=19

Yuan, K., & Le, V. N. (2014). Measuring deeper learning through cognitively demanding test items: Results from the analysis of six national and international exams. Research Report. *RAND Corporation*.

# Appendices

## Appendix A

Research Question 3 Survey



The UNIVERSITY of OKLAHOMA

**Default Question Block**

**Would you like to be involved in research at the University of Oklahoma?**

I am Leedy Smith, a doctoral student working under the supervision of Dr. Timothy G. Ford at the University of Oklahoma-Tulsa from the Department of Educational Leadership and Policy Studies, and I invite you to participate in my research project entitled Assessment systems and measures of deeper learning: An evaluation of USA TestPrep®. This research is being conducted at Bixby Middle School. You were selected as a possible participant because you may have used USA TestPrep® as part of your instructional duties with the district. You must be at least 18 years of age to participate in this study.

**Please read this document and contact me to ask any questions that you may have BEFORE agreeing to take part in my research.**

**What is the purpose of this research?** The purpose of this research is to gather evidence regarding the usefulness of USA TestPrep® for helping teachers achieve deeper learning outcomes for students and in turn prepare students who are college and career ready.

**How many participants will be in this research?** About 40 teachers will take part in this research.

**What will I be asked to do?** If you agree to be in this research, you will be asked to complete an online questionnaire regarding your experiences with and perceptions of USA TestPrep®.

**How long will this take?** Your participation will take about 10 minutes.

**What are the risks and/or benefits if I participate?** There are no benefits from being in this research. The questionnaire allows for negative comments and thus there is minimal employment risk if the research records are accidentally released. The likelihood of data being deductively re-identifiable is minimal. The likelihood of research records being accidentally released is minimal as well.

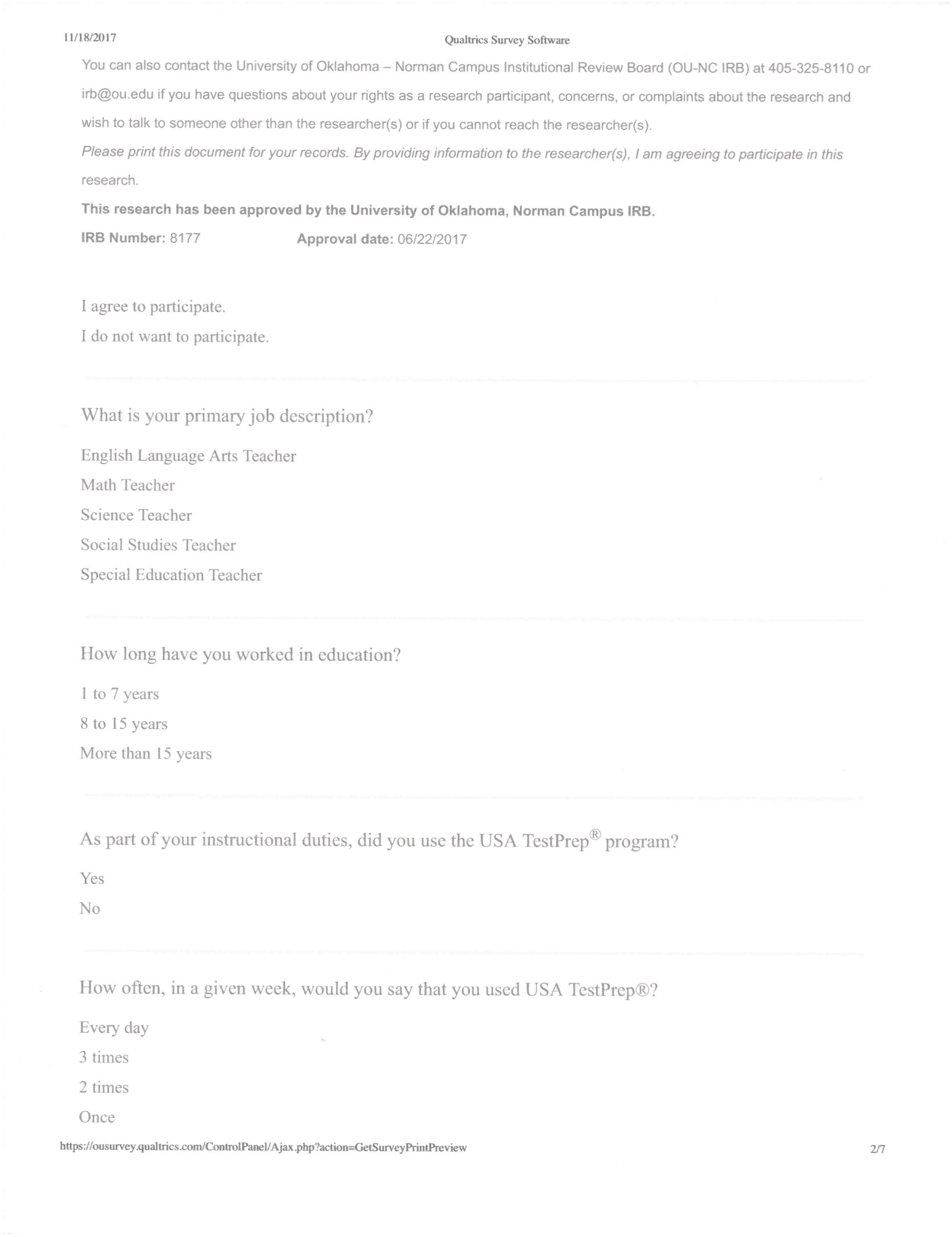
**Will I be compensated for participating?** You will not be reimbursed for your time and participation in this research.

**Who will see my information?** In research reports, there will be no information that will make it possible to identify you. Research records will be stored securely and only approved researchers and the OU Institutional Review Board will have access to the records.

Data are collected via an online survey system that has its own privacy and security policies for keeping your information confidential. Please note no assurance can be made as to the use of the data you provide for purposes other than this research.

**Do I have to participate?** No. If you do not participate, you will not be penalized or lose benefits or services unrelated to the research. If you decide to participate, you don't have to answer any question and can stop participating at any time.

**Who do I contact with questions, concerns or complaints?** If you have questions, concerns or complaints about the research or have experienced a research-related injury, contact me at 918-284-8339; Leedy.K.Smith-1@ou.edu or Dr. Timothy Ford at 918-660-3963; tgford@ou.edu.



11/18/2017 Qualtrics Survey Software

You can also contact the University of Oklahoma – Norman Campus Institutional Review Board (OU-NC IRB) at 405-325-8110 or irb@ou.edu if you have questions about your rights as a research participant, concerns, or complaints about the research and wish to talk to someone other than the researcher(s) or if you cannot reach the researcher(s).

*Please print this document for your records. By providing information to the researcher(s), I am agreeing to participate in this research.*

**This research has been approved by the University of Oklahoma, Norman Campus IRB.**

**IRB Number:** 8177 **Approval date:** 06/22/2017

I agree to participate.

I do not want to participate.

What is your primary job description?

English Language Arts Teacher

Math Teacher

Science Teacher

Social Studies Teacher

Special Education Teacher

How long have you worked in education?

1 to 7 years

8 to 15 years

More than 15 years

As part of your instructional duties, did you use the USA TestPrep® program?

Yes

No

How often, in a given week, would you say that you used USA TestPrep®?

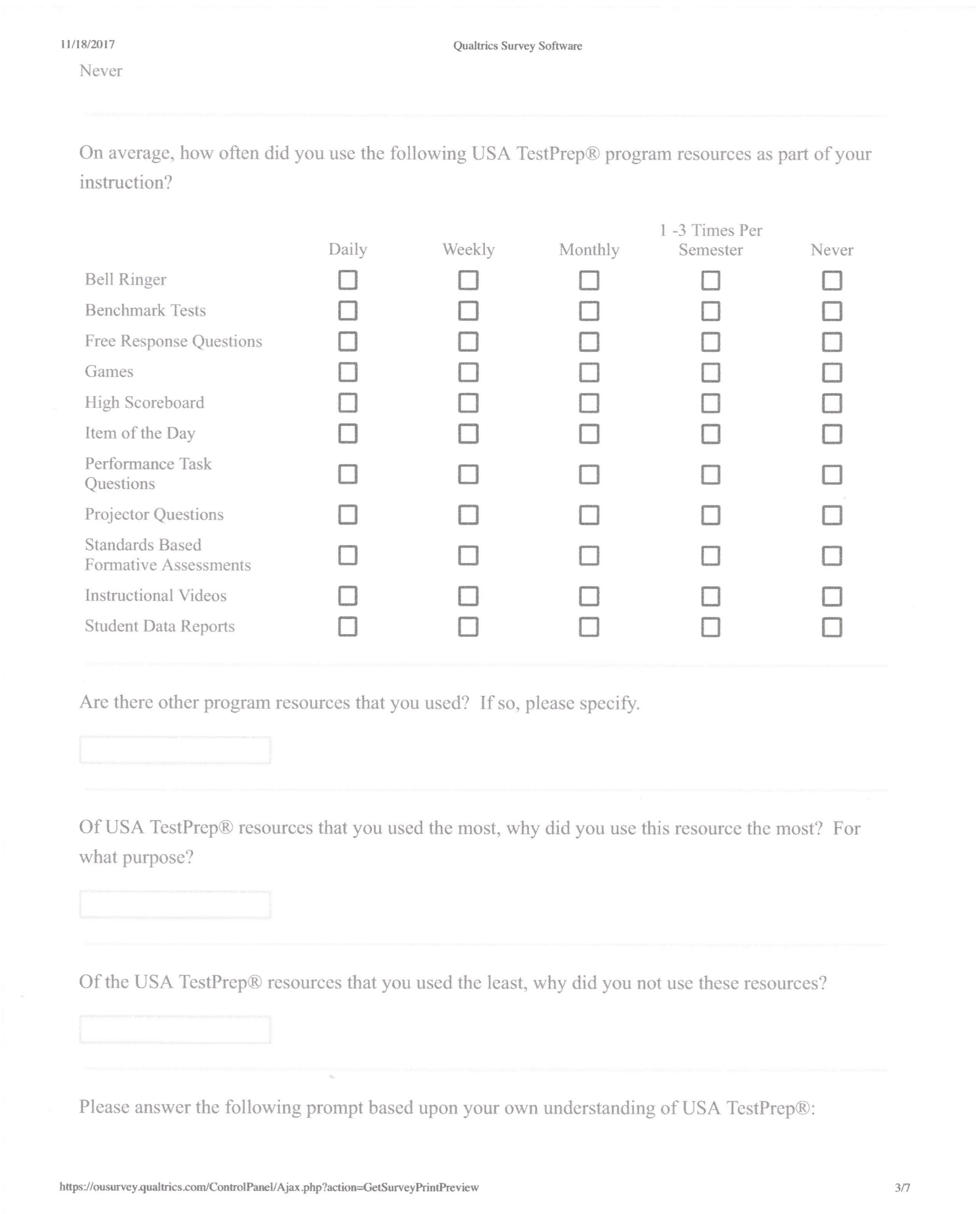
Every day

3 times

2 times

Once

https://ousurvey.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview 2/7



Never

On average, how often did you use the following USA TestPrep® program resources as part of your instruction?

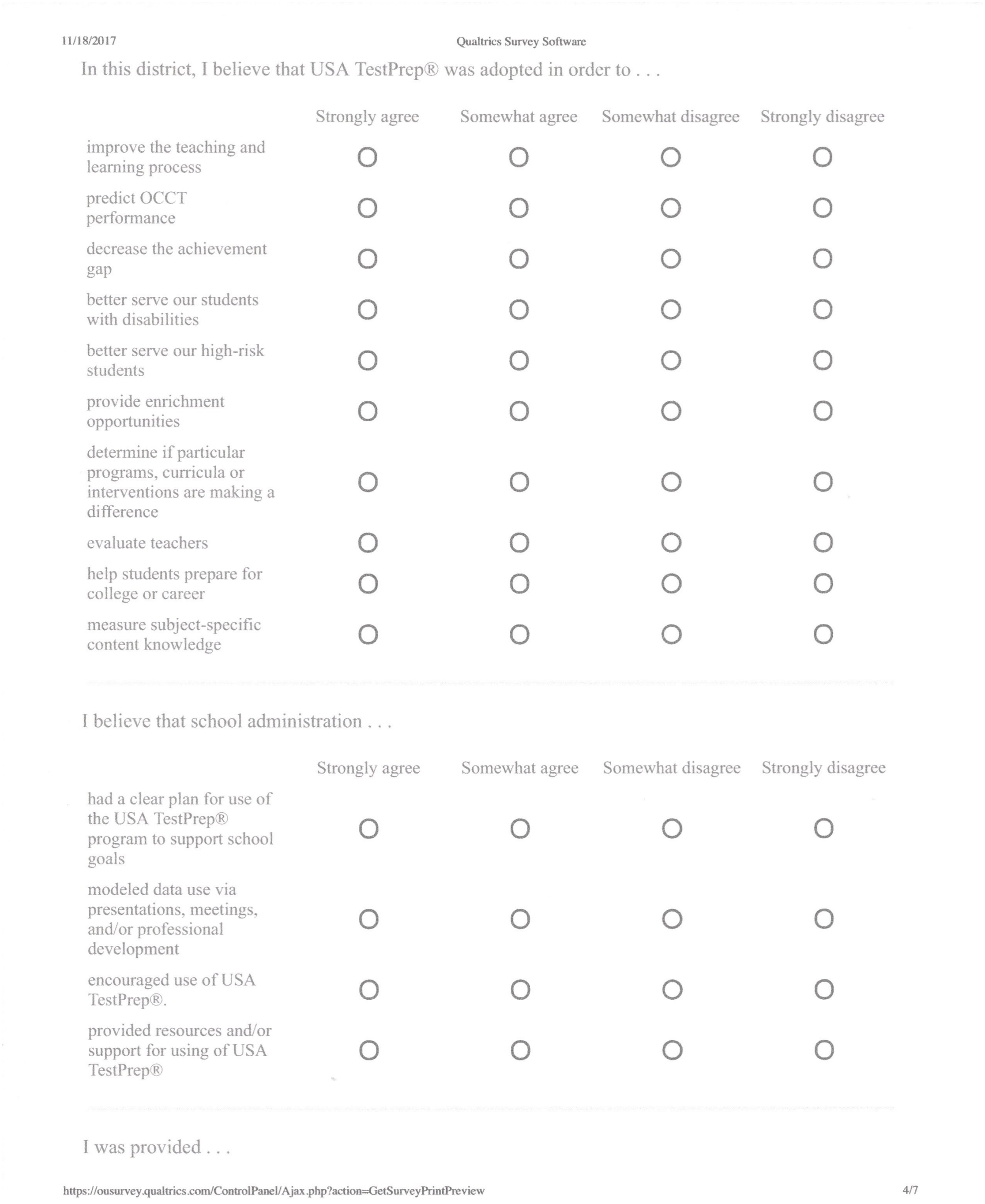
| | Daily | Weekly | Monthly | 1 -3 Times Per Semester | Never |
|---|---|---|---|---|---|
| Bell Ringer | □ | □ | □ | □ | □ |
| Benchmark Tests | □ | □ | □ | □ | □ |
| Free Response Questions | □ | □ | □ | □ | □ |
| Games | □ | □ | □ | □ | □ |
| High Scoreboard | □ | □ | □ | □ | □ |
| Item of the Day | □ | □ | □ | □ | □ |
| Performance Task Questions | □ | □ | □ | □ | □ |
| Projector Questions | □ | □ | □ | □ | □ |
| Standards Based Formative Assessments | □ | □ | □ | □ | □ |
| Instructional Videos | □ | □ | □ | □ | □ |
| Student Data Reports | □ | □ | □ | □ | □ |

Are there other program resources that you used? If so, please specify.

Of USA TestPrep® resources that you used the most, why did you use this resource the most? For what purpose?

Of the USA TestPrep® resources that you used the least, why did you not use these resources?

Please answer the following prompt based upon your own understanding of USA TestPrep®:



11/18/2017 Qualtrics Survey Software

In this district, I believe that USA TestPrep® was adopted in order to . . .

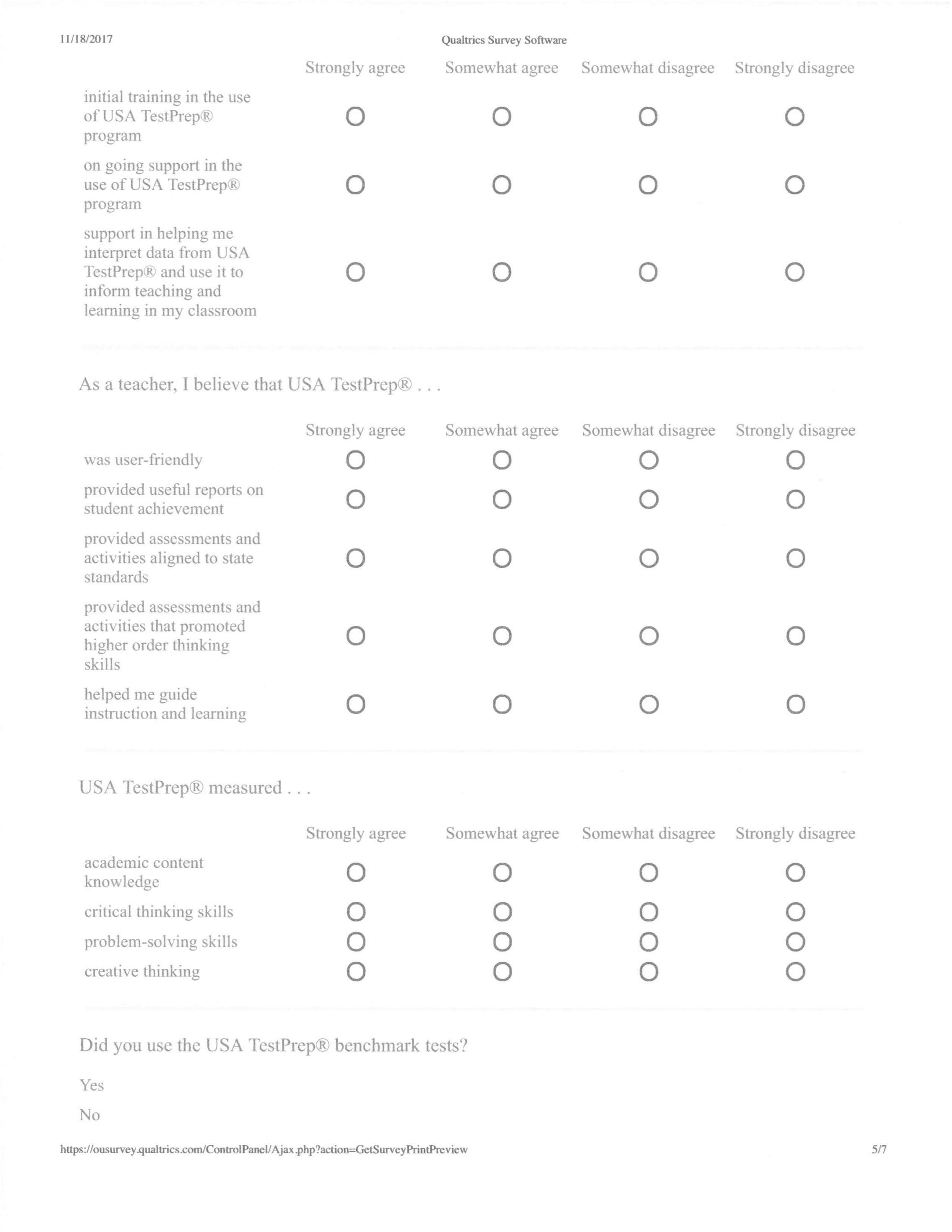
| | Strongly agree | Somewhat agree | Somewhat disagree | Strongly disagree |
|---|---|---|---|---|
| improve the teaching and learning process | ○ | ○ | ○ | ○ |
| predict OCCT performance | ○ | ○ | ○ | ○ |
| decrease the achievement gap | ○ | ○ | ○ | ○ |
| better serve our students with disabilities | ○ | ○ | ○ | ○ |
| better serve our high-risk students | ○ | ○ | ○ | ○ |
| provide enrichment opportunities | ○ | ○ | ○ | ○ |
| determine if particular programs, curricula or interventions are making a difference | ○ | ○ | ○ | ○ |
| evaluate teachers | ○ | ○ | ○ | ○ |
| help students prepare for college or career | ○ | ○ | ○ | ○ |
| measure subject-specific content knowledge | ○ | ○ | ○ | ○ |

I believe that school administration . . .

| | Strongly agree | Somewhat agree | Somewhat disagree | Strongly disagree |
|---|---|---|---|---|
| had a clear plan for use of the USA TestPrep® program to support school goals | ○ | ○ | ○ | ○ |
| modeled data use via presentations, meetings, and/or professional development | ○ | ○ | ○ | ○ |
| encouraged use of USA TestPrep®. | ○ | ○ | ○ | ○ |
| provided resources and/or support for using of USA TestPrep® | ○ | ○ | ○ | ○ |

I was provided . . .

https://ousurvey.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview 4/7



| | Strongly agree | Somewhat agree | Somewhat disagree | Strongly disagree |
|---|---|---|---|---|
| initial training in the use of USA TestPrep® program | ○ | ○ | ○ | ○ |
| on going support in the use of USA TestPrep® program | ○ | ○ | ○ | ○ |
| support in helping me interpret data from USA TestPrep® and use it to inform teaching and learning in my classroom | ○ | ○ | ○ | ○ |

As a teacher, I believe that USA TestPrep® . . .

| | Strongly agree | Somewhat agree | Somewhat disagree | Strongly disagree |
|---|---|---|---|---|
| was user-friendly | ○ | ○ | ○ | ○ |
| provided useful reports on student achievement | ○ | ○ | ○ | ○ |
| provided assessments and activities aligned to state standards | ○ | ○ | ○ | ○ |
| provided assessments and activities that promoted higher order thinking skills | ○ | ○ | ○ | ○ |
| helped me guide instruction and learning | ○ | ○ | ○ | ○ |

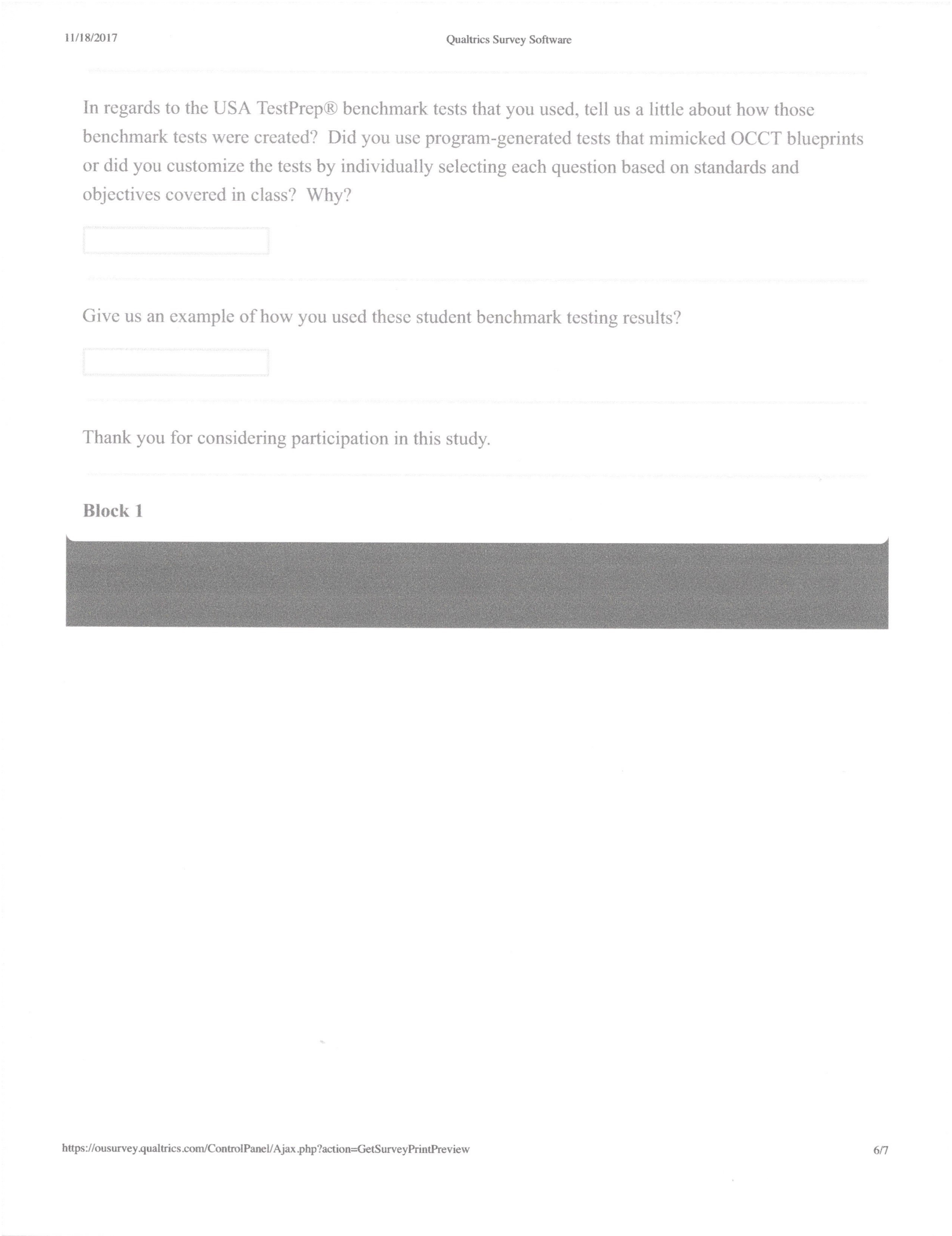
USA TestPrep® measured . . .

| | Strongly agree | Somewhat agree | Somewhat disagree | Strongly disagree |
|---|---|---|---|---|
| academic content knowledge | ○ | ○ | ○ | ○ |
| critical thinking skills | ○ | ○ | ○ | ○ |
| problem-solving skills | ○ | ○ | ○ | ○ |
| creative thinking | ○ | ○ | ○ | ○ |

Did you use the USA TestPrep® benchmark tests?

Yes

No



11/18/2017 Qualtrics Survey Software

In regards to the USA TestPrep® benchmark tests that you used, tell us a little about how those benchmark tests were created? Did you use program-generated tests that mimicked OCCT blueprints or did you customize the tests by individually selecting each question based on standards and objectives covered in class? Why?

Give us an example of how you used these student benchmark testing results?

Thank you for considering participation in this study.

**Block 1**

https://ousurvey.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview 6/7

# Appendix B

## Focus Group Protocol

**University of Oklahoma-Tulsa**

**Research Study: Interim Assessment Systems and Measures of Deeper Learning:**
**An Evaluation of USATestPrep®**
**Focus Group Questions**

Introductory Question:

1. Describe how you used the USA TestPrep® program?

Transition Question:

2. What is the first thing that comes to mind when you hear the phrase, "USA TestPrep®"?

Key Questions:

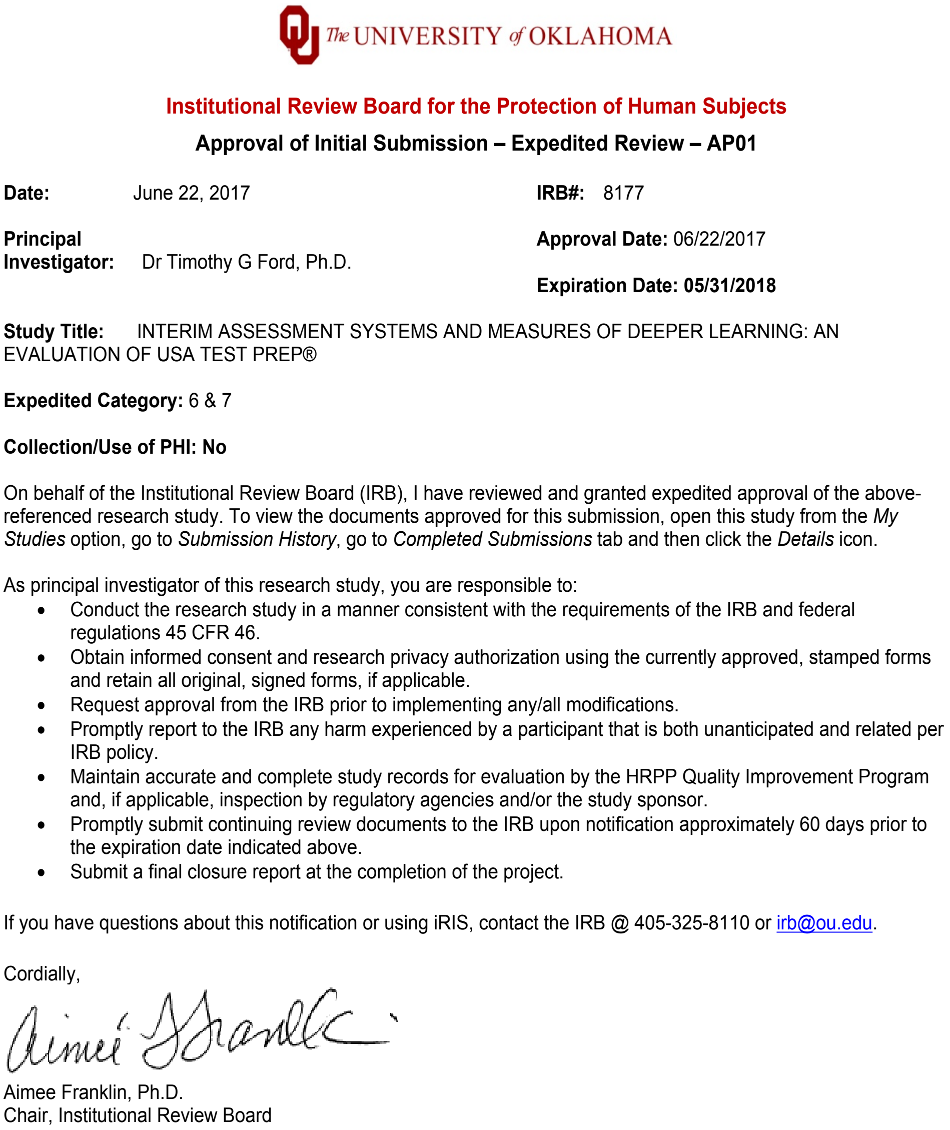
3. What did you find helpful about USA TestPrep®?
4. What did you find frustrating about USA TestPrep®?
5. How did USA TestPrep® influence learning and instruction in your classroom?

Ending Question:

6. Moderator will provide a brief summary of the conversation thus far and then ask, "How well does this capture what was said here today?
7. Is there anything else that you want to say about USA TestPrep® that you did not get a chance to say?

# Appendix C

IRB Approval Letter



The UNIVERSITY of OKLAHOMA

**Institutional Review Board for the Protection of Human Subjects**

**Approval of Initial Submission – Expedited Review – AP01**

|  |  |  |
|---|---|---|
| **Date:** June 22, 2017 | | **IRB#:** 8177 |
| **Principal Investigator:** Dr Timothy G Ford, Ph.D. | | **Approval Date:** 06/22/2017 |
| | | **Expiration Date: 05/31/2018** |

**Study Title:** INTERIM ASSESSMENT SYSTEMS AND MEASURES OF DEEPER LEARNING: AN EVALUATION OF USA TEST PREP®

**Expedited Category:** 6 & 7

**Collection/Use of PHI: No**

On behalf of the Institutional Review Board (IRB), I have reviewed and granted expedited approval of the above-referenced research study. To view the documents approved for this submission, open this study from the *My Studies* option, go to *Submission History*, go to *Completed Submissions* tab and then click the *Details* icon.

As principal investigator of this research study, you are responsible to:

- Conduct the research study in a manner consistent with the requirements of the IRB and federal regulations 45 CFR 46.
- Obtain informed consent and research privacy authorization using the currently approved, stamped forms and retain all original, signed forms, if applicable.
- Request approval from the IRB prior to implementing any/all modifications.
- Promptly report to the IRB any harm experienced by a participant that is both unanticipated and related per IRB policy.
- Maintain accurate and complete study records for evaluation by the HRPP Quality Improvement Program and, if applicable, inspection by regulatory agencies and/or the study sponsor.
- Promptly submit continuing review documents to the IRB upon notification approximately 60 days prior to the expiration date indicated above.
- Submit a final closure report at the completion of the project.

If you have questions about this notification or using iRIS, contact the IRB @ 405-325-8110 or irb@ou.edu.

Cordially,

Aimee Franklin

Aimee Franklin, Ph.D.
Chair, Institutional Review Board